International Society for Photogrammetry and Remote Sensing
Société Internationale de Photogrammétrie et de Télédétection
Internationale Gesellschaft für Photogrammetrie und Fernerkundung

THE INTERNATIONAL ARCHIVES OF THE PHOTOGRAMMETRY, REMOTE SENSING AND SPATIAL INFORMATION SCIENCES
ARCHIVES INTERNATIONALES DE PHOTOGRAMMÉTRIE, DE TÉLÉDÉTECTION ET DE SCIENCES DE L'INFORMATION SPATIALE
INTERNATIONALES ARCHIV FÜR PHOTOGRAMMETRIE, FERNERKUNDUNG UND RAUMBEZOGENE INFORMATIONSWISSENSCHAFTEN

VOLUME
VOLUME
BAND
XXXVIII

PART
TOME
TEIL
3 / W22

PIA11

Photogrammetric Image Analysis

Munich, Germany
October 5 – 7, 2011

Editors
U. Stilla, F. Rottensteiner, H. Mayer, B. Jutzi, M. Butenuth

Organised by
Department of Photogrammetry and Remote Sensing,
Technische Universitaet Muenchen (TUM)

in Cooperation with
ISPRS WG I/2 – LiDAR, SAR and Optical Sensors
ISPRS WG III/1 – Pose Estimation and Surface Reconstruction
ISPRS WG III/4 – Complex Scene Analysis and 3D Reconstruction
ISPRS WG III/5 – Image Sequence Analysis

ISSN 1682-1777

International Society for Photogrammetry and Remote Sensing
Société Internationale de Photogrammétrie et de Télédétection
Internationale Gesellschaft für Photogrammetrie und Fernerkundung

THE INTERNATIONAL ARCHIVES OF THE PHOTOGRAMMETRY, REMOTE SENSING AND SPATIAL INFORMATION SCIENCES
ARCHIVES INTERNATIONALES DE PHOTOGRAMMÉTRIE, DE TÉLÉDÉTECTION ET DE SCIENCES DE L'INFORMATION SPATIALE
INTERNATIONALES ARCHIV FÜR PHOTOGRAMMETRIE, FERNERKUNDUNG UND RAUMBEZOGENE INFORMATIONSWISSENSCHAFTEN

VOLUME
VOLUME
BAND
XXXVIII

PART
TOME
TEIL
3 / W22

PIA11

Photogrammetric Image Analysis

Munich, Germany
October 5 – 7, 2011

Editors
U. Stilla, F. Rottensteiner, H. Mayer, B. Jutzi, M. Butenuth

Organised by
Department of Photogrammetry and Remote Sensing,
Technische Universitaet Muenchen (TUM)

in Cooperation with
ISPRS WG I/2 – LiDAR, SAR and Optical Sensors
ISPRS WG III/1 – Pose Estimation and Surface Reconstruction
ISPRS WG III/4 – Complex Scene Analysis and 3D Reconstruction
ISPRS WG III/5 – Image Sequence Analysis

ISSN 1682-1777

Cooperating ISPRS Working Groups
- WG I/2 – LiDAR, SAR and Optical Sensors
- WG III/1 – Pose Estimation and Surface Reconstruction
- WG III/4 – Complex Scene Analysis and 3D Reconstruction
- WG III/5 – Image Sequence Analysis

ISPRS Headquarters 2008-2012
c/o CHEN JUN, ISPRS Secretary General
National Geomatics Center of China
28 Lianhuachixi Road, Haidian District
Beijing 100830, PR China
Tel: +86 10 6388 1102
Fax: +86 10 6388 1905
Email: chenjun@nsdi.gov.cn; chenjun_isprs@263.net
ISPRS WEB Homepage: http://www.isprs.org

Published by
Department of Photogrammetry and Remote Sensing
Technische Universitaet Muenchen (TUM)

Available from
GITC bv
P.O.Box 112
8530 AC Lemmer
The Netherlands
Tel: +31 (0) 514 56 18 54
Fax: +31 (0) 514 56 38 98
E-mail: mailbox@gitc.nl
Website: www.gitc.nl

Conference Committees

PIA11 Conference Chair and Co-Chairs:

Uwe Stilla, Technische Universitaet Muenchen (TUM), Germany
Franz Rottensteiner, Leibniz Universitaet Hannover, Germany
Helmut Mayer, Universitaet der Bundeswehr Muenchen, Germany
Boris Jutzi, Karlsruhe Institute of Technology (KIT), Germany
Matthias Butenuth, Technische Universitaet Muenchen (TUM), Germany

PIA11 Program Committee

Michael Arens, Fraunhofer IOSB, Germany
Caroline Baillard, SIRADEL, France
Richard Bamler, German Aerospace Center (DLR), Germany
Matthias Butenuth, Technische Universitaet Muenchen (TUM), Germany
Ismael Colomina, Institut de Geomatica Castelldefels, Spain
Wolfgang Foerstner, University of Bonn, Germany
Jan-Michael Frahm, University of North Carolina, USA
Markus Gerke, University of Twente, Netherlands
Norbert Haala, University of Stuttgart, Germany
Christian Heipke, Leibniz Universitaet Hannover, Germany
Olaf Hellwich, Technische Universitaet Berlin, Germany
Stefan Hinz, Karlsruhe Institute of Technology (KIT), Germany
Boris Jutzi, Karlsruhe Institute of Technology (KIT), Germany
Clement Mallet, Institut Geographique National (IGN), France
Helmut Mayer, Universitaet der Bundeswehr Muenchen, Germany
Chris McGlone, SAIC, USA
Jochen Meidow, Fraunhofer IOSB, Germany
Franz Josef Meyer, University of Alaska Fairbanks, USA
Stephan Nebiker, University of Applied Sciences Northwestern Switzerland, Switzerland
Nicolas Paparoditis, Institut Geographique National (IGN), France
Camillo Ressl, Vienne University of Technology, Austria
Franz Rottensteiner, Leibniz Universitaet Hannover, Germany
Konrad Schindler, ETH Zuerich, Switzerland
Uwe Soergel, Leibniz Universitaet Hannover, Germany
Gunho Sohn, York University, USA
Uwe Stilla, Technische Universitaet Muenchen (TUM), Germany
Christoph Strecha, EPLF, Switzerland
Charles Toth, Ohio State University, USA
Yongjun Zhang, Wuhan Univiersity, China

PIA11 Local Organizing Committee

Florian Burkert, Technische Universitaet Muenchen (TUM), Germany
Konrad Eder, Technische Universitaet Muenchen (TUM), Germany
Christine Elmauer, Technische Universitaet Muenchen (TUM), Germany
Carsten Goetz, Technische Universitaet Muenchen (TUM), Germany
Ludwig Hoegner, Technische Universitaet Muenchen (TUM), Germany
Dorota Iwaszczuk, Technische Universitaet Muenchen (TUM), Germany
Michael Schmitt, Technische Universitaet Muenchen (TUM), Germany
Sebastian Tuttas, Technische Universitaet Muenchen (TUM), Germany

Preface

Automated extraction of objects from remotely sensed data is an important topic of research in Photogrammetry, Computer Vision, Remote Sensing, and Geoinformation Science. PIA11 addressed researchers and practitioners from universities, research institutes, industry, government organizations, and private companies. The range of topics covered by the conference is reflected by the terms of reference of the cooperating working groups of the International Society for Photogrammetry and Remote Sensing (ISRPS):

- Lidar, SAR and Optical Sensors (WG I/2)
- Pose Estimation and Surface Reconstruction (WG III/1)
- Complex Scene Analysis and 3D Reconstruction (WG III/4)
- Image Sequence Analysis (WG III/5)

After the successful series of ISPRS conferences on Photogrammetric Image Analysis in Munich in 1999, 2003, and 2007, in 2011 PIA11 again discussed recent developments, the potential of various data sources, and future trends in automated object extraction with respect to both sensors and processing techniques, focusing on methodological research. It was held at Technische Universitaet Muenchen (TUM) in Munich, Germany, 5-7 October 2011.

Prospective authors were invited to submit full papers of a maximum length of six A4 pages. We received 54 full papers coming from 18 countries for review. The submitted papers were subject to a rigorous double blind peer review process. Forty-two papers were reviewed by three members of the program committee, whereas the rest (12 papers) was reviewed by two members of that committee. In total we received 150 reviews from 29 reviewers. Altogether 30 papers were accepted based on the reviews, which corresponds to a rejection rate of 44%. From those 25 papers were published in printed form within the book series 'Lecture Notes in Computer Science' (LNCS) of Springer-Verlag and 5 papers are contained in this volume of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. All contributions are listed in Part 1. Additionally, authors who intended to present application-oriented work particularly suitable for interactive presentation were invited to submit extended abstracts. Part 2 of this volume contains 24 of these papers.

Altogether, PIA11 featured 7 oral sessions, 2 poster sessions, and 2 invited talks, namely "Convex optimization methods for Computer Vision" (Daniel Cremers) and "Exploiting redundancy for reliable aerial Computer Vision" (Horst Bischof).

Finally, the editors wish to thank all contributing authors and the members of the Program Committee. In addition, we would like to express our thanks to the Local Organizing Committee, without whom this event could not have taken place. Ludwig Hoegner did a great job managing of the conference tool. The final editing of all incoming manuscripts and the preparation of the proceedings by Michael Schmitt are gratefully acknowledged. Konrad Eder and Dorota Iwaszczuk did a great job organizing the social events and accomodation, Florian Burkert in caring for the technical equipment, and Sebastian Tuttas in supervising the local organizing committee assistants. We would also like to thank Christine Elmauer, Carsten Goetz, and Gabriele Aumann for their support to make PIA11 a successful event.

Last, but not least we would like to thank our sponsors MVTec Software GmbH and INPHO GmbH – A TRIMBLE COMPANY, and our supporting institutions ISPRS, ASPRS, DGPF, EuroSDR, EARSeL and IAG for their assistance.

Munich, October 2011

The conference chairs





U. Stilla F. Rottensteiner H. Mayer B. Jutzi M. Butenuth

Contents

Part 1: Papers accepted by full paper review

Orientation

Matching

Object Detection

3D-Reconstruction and DEM

Classification

People and Tracking

Image Processing and Visualization

Part 2: Papers accepted by extended abstract review

Orientation

Matching and Registration

Object Detection

3D-Reconstruction and DEM

Classification

Vehicles and People

Part 1

Papers accepted by full paper review

EFFICIENT VIDEO MOSAICKING BY MULTIPLE LOOP CLOSING

J. Meidow

Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, 76275 Ettlingen, Germany – jochen.meidow@iosb.fraunhofer.de

Working Groups I/2, III/1, III/4, III/5

KEYWORDS: image alignment, mosaic, loop closing, homography, exponential representation, parameter estimation

ABSTRACT:

The rapid generation of aerial mosaics is an important task for change detection, e.g. in the context of disaster management or surveillance. Unmanned aerial vehicles equipped with a single camera offer the possibility to solve this task with moderate efforts. Unfortunately, the accumulation of tracking errors leads to a drift in the alignment of images which has to be compensated by loop closing for instance. We propose a novel approach for constructing large, consistent and undistorted mosaics by aligning video images of planar scenes. The approach allows the simultaneous closing of multiple loops possibly resulting from the camera path in a batch process. The choice of the adjustment model leads to statistical rigorous solutions while the used minimal representations for the involved homographies and the exploitation of the natural image order enable very efficient computations. The approach will be empirically evaluated with the help of synthetic data and its feasibility will be demonstrated with real data sets.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors:	Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume:	6952
Series Editors:	Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN:	0302-9743

The article is accessible online through www.springerlink.com.

ESTIMATING THE MUTUAL ORIENTATION IN A MULTI-CAMERA SYSTEM WITH A NON OVERLAPPING FIELD OF VIEW

D. Muhle[1], S. Abraham[2], C. Heipke[1], M. Wiggenhagen[1]

[1] Institut für Photogrammetrie und GeoInformation, Nienburgerstr. 1, 30167 Hannover, Germany
[2] Robert Bosch GmbH, Robert-Bosch-Str. 200, 31139 Hildesheim

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: mutual orientation, non-overlapping field of view, bundle adjustment, motion, image sequence

ABSTRACT:

Multi-camera systems offer some advantages over classical systems like stereo or monocular camera systems. A multi-camera system with a non-overlapping field of view, able to cover a wide area, might prove superior e. g. in a mapping scenario where less time is needed to cover the entire area. Approaches to determine the parameters of the mutual orientation from common motions exist for more than 30 years. Most work presented in the past neglected or ignored the influence different motion characteristics have on the parameter estimation process. However, for critical motions a subset of the parameters of the mutual orientation can not be determined or only very inaccurate. In this paper we present a strategy and assessment scheme to allow a successful estimation of as many parameters as possible even for critical motions. Furthermore, the proposed approach is validated by experiments.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

ABSOLUTE ORIENTATION OF STEREOSCOPIC CAMERAS BY ALIGNING CONTOURS IN PAIRS OF IMAGES AND REFERENCE IMAGES

B.P. Selby[1], G. Sakas[1], W.-D. Groch[2], U. Stilla[3]

[1] Medcom GmbH, Image Guided RT, Darmstadt, Germany
[2] Dept. of Computer Sciences, University of Applied Sciences, Darmstadt, Germany
[3] Dept. of Photogrammetry and Remote Sensing, Technische Universitaet Muenchen, Muenchen, Germany

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: camera orientation, contour registration, automatic alignment

ABSTRACT:

Most approaches use corresponding points to determining an object's orientation from stereo-images, but this is not always possible. Imaging modalities that do not produce correspondences for different viewing angles, as in X-ray imaging, require other procedures. Our method works on contours in images that do not need to be equivalent in length or contain corresponding points. It is able to determine corresponding contours and resamples those, creating new sets of corresponding points for registration. Two sets of in-plane transformations from a stereo-system are used to determine spatial orientation. The approach was tested with three ground truth datasets and sub-pixel accuracy was achieved. The approach is originally designed for X-ray based patient alignment, but it is versatile and can also be employed in other close range photogrammetry applications.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

MATCHING BETWEEN DIFFERENT IMAGE DOMAINS

C. Toth, H. Ju, D. Grejner-Brzezinska

The Center for Mapping, Ohio State University, 470 Hitchcock Hall, 2070 Neil Avenue, Columbus, OH 43210 – toth@cfm.ohio-state.edu

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: image registration/matching, LiDAR, satellite imagery

ABSTRACT:

Most of the image registration/matching methods are applicable to images acquired by either identical or similar sensors from various positions. Simpler techniques assume some object space relationship between sensor reference points, such as near parallel image planes, certain overlap and comparable radiometric characteristics. More robust methods allow for larger variations in image orientation and texture, such as the Scale-Invariant Feature Transformation (SIFT), a highly robust technique widely used in computer vision. The use of SIFT, however, is quite limited in mapping so far, mainly, because most of the imagery are acquired from airborne/spaceborne platforms, and, consequently, the image orientation is better known, presenting a less general case for matching. The motivation for this study is to look at the feasibility of a particular case of matching between different image domains. In this investigation, the co-registration of satellite imagery and LiDAR intensity data is addressed.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

RELIABLE IMAGE MATCHING WITH RECURSIVE TILING

D. Novak, E. Baltsavias, K. Schindler

Institute of Geodesy and Photogrammetry, ETH Zürich, 8093 Zürich, Switzerland

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: matching, processing, orientation, reconstruction

ABSTRACT:

This paper presents a method to improve the robustness of automated correspondences while also increasing the total amount of measured points and improving the point distribution. This is achieved by incorporating a tiling technique into existing automated interest point extraction and matching algorithms. The technique allows memory intensive interest point extractors like SIFT to use large images beyond 10 megapixels while also making it possible to approximately compensate for perspective differences and thus get matches in places where normal techniques usually do not get any, few, or false ones. The experiments in this paper show an increased amount as well as a more homogeneous distribution of matches compared to standard procedures.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

IN-STRIP MATCHING AND RECONSTRUCTION OF LINE SEGMENTS FROM UHR AERIAL IMAGE TRIPLETS

A.Ö. Ok[1], J.D. Wegner[2], C. Heipke[2], F. Rottensteiner[2], U. Soergel[2], Vedat Toprak[1]

[1] Dept. of Geodetic and Geographic Information Tech., Middle East Technical University, 06531 Ankara, Turkey – (oozgun, toprak)@metu.edu.tr
[2] Institute of Photogrammetry and Geoinformation, University of Hannover, 30167 Hannover, Germany – (wegner, heipke, rottensteiner, soergel)@ipi.uni-hannover.de

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: three-view matching, in-strip line matching, reconstruction, aerial images, pair-wise matching

ABSTRACT:

In this study, we propose a new line matching and reconstruction methodology for aerial image triplets that are acquired within a single strip. The newly developed stereo reconstruction approach gives us better line predictions in the third image which in turn helps to improve the performance of the matching. The redundancy information generated in each stereo match gives us ability to reduce the number of false matches while preserving high levels of matching completeness. The approach is tested over four test patches and produced highly promising line matching and reconstruction results.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

REFINED NON-RIGID REGISTRATION OF A PANORAMIC IMAGE SEQUENCE TO A LIDAR POINT CLOUD

A. Swart[1,2], J. Broere[1], R. Veltkamp[2], R. Tan[2]

[1] Cyclomedia Technology BV, Waardenburg, Netherlands –
(aswart, jbroere)@cyclomedia.com
[2] Utrecht University, Utrecht, Netherlands –
(remco.veltkamp,robby)@cs.uu.nl

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: mobile mapping, LiDAR, multi sensor, registration, point cloud, panorama

ABSTRACT:

The combination of LiDAR data with panoramic images could be of great benefit to many geo-related applications and processes such as measuring and map making. Although it is possible to record both LiDAR points and panoramic images at the same time, there are economic and practical advantages to separating the acquisition of both types of data. However, when LiDAR and image data is recorded separately, poor GPS reception in many urban areas will make registration between the data sets necessary. In this paper, we describe a method to register a sequence of panoramic images to a LiDAR point cloud using a non-rigid version of ICP that incorporates a bundle adjustment framework. The registration is then refined by integrating image-to-reflectance data SIFT correspondences into the bundle adjustment. We demonstrate the validity of this registration method by a comparison against ground truth data.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

IMAGE SEQUENCE PROCESSING IN STEREOVISION MOBILE MAPPING – STEPS TOWARDS ROBUST AND ACCURATE MONOSCOPIC 3D MEASUREMENTS AND IMAGE-BASED GEOREFERENCING

F. Huber[1], S. Nebiker[1], H. Eugster[1,2]

[1] FHNW, University of Applied Sciences Northwestern Switzerland, School of Architecture, Civil Engineering and Geomatics, Inst. of Geomatics Engineering, CH-4132 Muttenz, Switzerland – (fabian.huber, stephan.nebiker, hannes.eugster)@fhnw.ch
[2] iNovitas AG, Mobile Mapping Solutions, Gründenstrasse 40, CH-4132 Muttenz, Switzerland – hannes.eugster@inovitas.ch

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: stereoscopic vision, mobile mapping, image matching, sequence processing

ABSTRACT:

Stereo vision based mobile mapping systems enable the efficient capturing of directly georeferenced stereo pairs. With today's camera and storage technologies imagery can be captured at high data rates resulting in dense stereo sequences. The overlap within stereo pairs and stereo sequences can be exploited to improve the accuracy and reliability of point measurements. This paper aims at robust and accurate monoscopic 3d measurements in future vision-based mobile mapping services. Key element is an adapted Least Squares Matching approach yielding point matching accuracies at the subpixel level. Initial positions for the matching process along the stereo sequence, are obtained by projecting the matched point position within the reference stereo pair to object space and by reprojecting it to the adjacent pairs. Once homologue image positions have been derived, final 3D point coordinates are estimated. Investigations with real-world data show, that points can successfully and reliably be matched over extended stereo sequences.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

GABLE ROOF DETECTION IN TERRESTRIAL IMAGES

V. Brandou, C. Baillard

SIRADEL, 3 allée Adolphe Bobierre, CS 14343, 35043 Rennes, France –
(vbrandou, cbaillard)@siradel.com

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: terrestrial, building, detection, classification, cartography, modelling

ABSTRACT:

This paper presents an automatic method for gable roof detection in terrestrial images. The purpose of this study is to refine the roofs of a 3D city model automatically derived from aerial images. The input images consist of geo-referenced terrestrial images acquired by a mobile mapping system (MMS). The raw images have been rectified and merged into seamless façade texture images (one texture per façade). Firstly, each image is pre-processed in order to remove small structures and to smooth homogeneous areas. Secondly, line segments are extracted and analysed to define the lateral edges of the roof. Finally, the analysis of the lowest part of the roof leads to the classification of the roof as gable or non-gable. The method was tested on more than 150 images and shows promising results.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

MULTI-SPECTRAL FALSE COLOR SHADOW DETECTION

M. Teke[1], E. Baseski[1], A.Ö. Ok[2], B. Yüksel[1], C. Senaras[1]

[1] HAVELSAN A.S., Eskisehir Yolu 7.km 06520, Ankara, Turkey –
(mteke, ebaseski, byuksel, csenaras)@havelsan.com.tr
[2] Middle East Technical University, Department of Geodetic and Geographic Information Technologies, 06531, Ankara, Turkey –
oozgun@metu.edu.tr

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: shadow detection, building detection, near-infrared, false color

ABSTRACT:

With the availability of high-resolution commercial satellite images, automated analysis and object extraction became even a more important topic in remote sensing. As shadows cover a significant portion of an image, they play an important role on automated analysis. While they degrade performance of applications such as image registration, shadow is an important cue for information such as man-made structures. In this article, a shadow detection algorithm that makes use of near-infrared information in combination with RGB bands is introduced. The algorithm is applied on an application for automated building detection.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

EXTRACTION OF NON-FOREST TREES FOR BIOMASS ASSESSMENT BASED ON AIRBORNE AND TERRESTRIAL LIDAR DATA

M. Rentsch[1], A. Krismann[2], P. Krzystek[1]

[1] Munich University of Applied Sciences, Department of Geoinformatics, Karlstr. 6, 80333 Munich, Germany – (rentsch, krzystek)@hm.edu
[2] University of Hohenheim, Institute for Landscape- and Plant-Ecology, August-v.-Hartmann-Str. 3, 70593 Stuttgart, Germany – a_krismann@uni-hohenheim.de

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: LiDAR, vegetation, correlation, point cloud, segmentation, three-dimensional

ABSTRACT:

The main goal of the federal funded project 'LiDAR based biomass assessment' is the nationwide investigation of the biomass potential coming from wood cuttings of non-forest trees. In this context, first and last pulse airborne laserscanning (F+L) data serve as preferred database. First of all, mandatory field calibrations are performed for pre-defined grove types. For this purpose, selected reference groves are captured by full-waveform airborne laserscanning (FWF) and terrestrial laserscanning (TLS) data in different foliage conditions. The paper is reporting about two methods for the biomass assessment of non-forest trees. The first method covers the determination of volume-to-biomass conversion factors which relate the reference above-ground biomass (AGB) estimated from allometric functions with the laserscanning derived vegetation volume. The second method is focused on a 3D Normalized Cut segmentation adopted for non-forest trees and the follow-on biomass calculation based on segmentation-derived tree features.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

CHANGE DETECTION IN A TOPOGRAPHIC BUILDING DATABASE USING SUBMETRIC SATELLITE IMAGES

Arnaud Le Bris [a] and Nesrine Chehata [a,b]

[a] Université Paris-Est, IGN, Laboratoire MATIS
73 Avenue de Paris, 94165 SAINT-MANDE Cedex, FRANCE
arnaud.le-bris@ign.fr
http://recherche.ign.fr/labos/matis/
[b] G&E Laboratory, ENSEGID, Bordeaux University
1 Allée F. Daguin, 33607 PESSAC Cedex, FRANCE
nesrine.chehata@egid.u-bordeaux3.fr

KEY WORDS: Image classification - Land cover extraction - Data fusion - Change detection - Satellite - DSM

ABSTRACT:

Submetric satellite imagery (Pleiades, GeoEye) offers advantages for map update purposes, e.g. an interesting ground resolution, a good reactivity and the ability to capture wide areas. Experiments on the use of such stereoscopic images for 2D change detection among building objects of GIS topographic database are presented in this paper. Two approaches have been tested. The first one extracts land cover from satellite ortho-images and additional information (correlation DSM-DTM, database) and compares building objects of this classification to those of the database. The second one creates a pseudo-DSM from height information of database building objects combined with a DTM and compares it to a correlation DSM computed from satellite images. Obtained results are quite encouraging even if the correctness rate remains too low for an operational use.

1 INTRODUCTION

Nowadays, most mapping agencies have finished the initial plotting of their topographic (2D or 2.5D) databases. Therefore, updating methods have become an important issue. Manual change detection is indeed costly and very long for agencies. Besides, it is a difficult and quite boring task for operators. In some mapping agencies, processes have been settled to catch information about changes from local authorities, giving good results for some themes such as roads, but being sometimes not sufficient for other ones such as buildings, especially in very changing areas. Therefore, there is a growing need for (semi-)automatic tools launching alarms (where change hypotheses are detected) and sending them to be checked by an operator. Such tools should be very exhaustive and as correct as possible (i.e. minimizing the false detection rate). In this paper, change detection is focused on **building theme** and only **2D changes** are sought (even if 3D information is used).
Change detection has been studied for years using various sensors and testing different approaches. LIDAR data associated with very high resolution (10-20cm) aerial images is often used as input data ((Rottensteiner, 2007, Matikainen et al., 2010)). (Poulain et al., 2009) use radar data associated with optic imagery. In the present paper, submetric satellite imagery (Pleiades, GeoEye) is used. Such sensors offer indeed advantages for database updating purposes, since they have an interesting ground resolution, a good reactivity and the ability to capture wide areas. Nevertheless, such data is less easy to use than very high resolution lidar data or aerial images : some smaller details can indeed be missed.
Among various approaches presented in literature, some consisting in comparing new images to old ones can be cited (Radke et al., 2005), but seem difficult to use in this special context. Many other approaches compare DSMs calculated at both dates. Other ones extract primitives from new data and compare them to the database to update ((Poulain et al., 2009, Champion et al., 2010)). (Bouziani et al., 2010) also use context information.
The goal here is to obtain alarms on changed building parts, to be checked by operators. Therefore, as the point is to detect new or demolished pixels, there is no need to discriminate between new, demolished and modified buildings (since a modified, bad plotted or misregistered building has "new" or "demolished" parts...). Two distinct approaches are used.
The first one extracts land cover from satellite (ortho-)images and additional information (database, DSM-DTM when available...) owing to a supervised per region classification process trained by database buildings and road objects. Unlike (Rottensteiner, 2007, Olsen and Knudsen, 2005), a radiometric model is trained from data and not only *ndvi* is used as radiometric information. Some buildings have indeed specific colours (e.g. white or bright red) not found in other classes, even if building theme remains difficult to model because of the variety of roofing materials. Unlike (Matikainen et al., 2010) (who uses lidar data), segmentation is computed from image and not from DSM, because DSMs calculated from satellite stereoscopic images often remain quite noisy with not well delineated object borders, whereas image offers more accurate contours.
The second approach is based on interdate DSMs comparison. The DSM generated from the new satellites images is compared to a "pseudo-DSM" calculated from a DTM and height information of database building objects. Finally alarms obtained from these two methods are merged.

2 DATA SETS AND MATERIALS

Experiments were carried out on two test zones located near Toulouse (France) and an other zone in La Réunion, with different data.

2.1 Toulouse "CHU" and "highway" areas

"CHU" area is a mixed urban - rural area, covering almost 1.5 km^2. A hill is present. Different kinds of buildings are present on this area, with big buildings (a school and a hospital) and individual houses (almost 230 buildings). Vegetation is quite important too, consisting of woods and fields.

"Highway" area is an urban area, covering almost 1 km^2. The ground is quite flat. It is mostly a residential suburb consisting of houses (almost 1200 buildings). Several more large buildings are present too. Road network is important, with a motorway, a cloverleaf intersection and bridges. There is less vegetation than in "CHU" test area. This is a "urban" vegetation consisting of trees (on the border of the streets or in gardens) and of lawns (in gardens, stadiums and squares).

Simulated Pleiades imagery (CNES, 2011) Tri-stereoscopic four bands (red - green - blue - near infrared) Pleiades simulated images are available for these regions. A 50cm ground resolution DSM has been calculated from the Pleiades images thanks to the automatic correlation tool described in (Pierrot-Deseilligny and Paparoditis, 2006). 50cm resolution true ortho-images have then been derived from the simulated Pleiades images and are used in the process.

"Reference" vector database is a detailed building and road topographic database plotted for Pleiades studies, with one object per building, and it also offers an accurate height information for building objects.

A **DTM** from IGN's national DTM database is used. It is a 25m ground resolution one.

2.2 "La Réunion" test area

This is also a mixed urban - rural area, covering almost 1.5 km^2. Relief is stronger there. Different kinds of buildings are present, with "industrial" buildings and individual houses. Nevertheless, these houses are often low (only one level) and covered by more kinds of roofing materials (different colours) making classification task more difficult than for Toulouse areas. Vegetation is quite important too, consisting of woods, fields and gardens.

GeoEye-1 imagery (GeoEye, 2011) A stereoscopic pair of GeoEye-1 four bands (red-green-blue-near infrared) images is used. Both images were captured at one week apart. As for previous test zones, a correlation DSM and a 50cm resolution true ortho-image have been calculated.

IGN's "BDTopo" topographic vector database is the database to update here. It contains road and building themes, but building objects of this national reference database are more generalized than the ones of the "Reference" database used on Toulouse test areas. Moreover, a BDTopo building object may contain several individual buildings, especially in dense urban areas. This database also offers a height information for building objects. 791 building objects are present in this test area.

The used **DTM** has been generated from the correlation DSM using "elastic grid" tools described in (Champion et al., 2010).

3 DIFFICULTIES AND PROPOSED SOLUTIONS

Several difficulties are encountered. Some of them are classic ones related to land cover classification (as for instance areas belonging to distinct classes but looking like each other, shadows...). Other limits are more specific to the approach and to its change detection purpose (as the presence of partly hidden objects, such as road section masked by trees or shadows).

3.1 Distinct classes with similar radiometry

Some distinct classes may have very similar radiometric distributions, making them very hard to be distinguished using only radiometric information.

Roads and grey roofed buildings (fig. 1) have almost the same radiometric distribution for the four bands (red - green - blue - near infrared) of the image. As a consequence, no derived channels from original image bands can really improve the discrimination between both classes. Unfortunately, the two themes "buildings" and "roads" are the most important themes to update in the database and have therefore not to be misclassified.

Buildings covered by red tiles are sometimes misclassified with **bare soil** belonging to fields or paths (fig. 1). These errors are a problem too, since "building" theme (which has to be updated) is concerned.

Figure 1: Radiometric information is not sufficient to distinguish classes with similar radiometry, such as grey roofed buildings from roads, or red tiles from bare soil.

Low and high vegetation areas are sometimes difficult to distinguish from each other. Nevertheless, this is not a problem here since these classes don't belong to the database to update.

These possible misclassifications always concern high and low objects (buildings / roads ; buildings / bare soil ; high vegetation / low vegetation). As a consequence, using information about the distinction between ground and above-ground areas is a possible solution. In case of a stereoscopic image acquisition, such knowledge can be derived from difference between DTM and correlation DSM and can be introduced into the classification process. To some extent, shadows could also be used to obtain such information.

3.2 Mixed classes and intra-class radiometric variations

Building objects can appear very different from each other on the image, since they are not covered by the same roofing material. Thus, they can not be described by a common radiometric model. Therefore, "building" class has to be considered as a mixed class and divided into "red roofed", "grey roofed" and "white roofed" building subclasses. At the end of the classification, these 3 classes will be merged into a single class "building".

Moreover, radiometry can greatly vary inside one class because of several factors such as the link between the orientation of an object and the sun illumination angle for "building" class, the density of cars or road marks in the streets for the theme "road". As in 3.1, introducing additional knowledge (such as belonging to ground/above ground or information derived from the old database) in the classification process can help.

3.3 Higher elements and shadows

Objects can be partly masked on the image by elements belonging to other classes. For instance, roads are often masked by shadows or higher objects such as trees (see fig. 2).
Shadows are obviously important in urban areas. They mostly concern streets but roofs can also be partly masked by shadows caused by higher buildings or roof superstructures. Therefore, an additional class "shadow" is also defined in order to take shadows into account.

These two phenomenons have two consequences here :
First, classification results can be "exact" (according to the ortho-image) whereas ground truth land cover is false, leading to false change detection.
Second, database objects used as training data can be masked and thus provide false training set. Nevertheless, since the database is not up-to-date, false training data will necessarily be provided in our case... A solution to cope with this consists in a two-pass process, calculating a first model and a first classification from all training data, and then a second model and a second classification from training data cleaned using the results of this first classification.

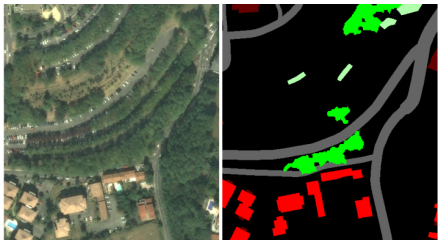


Figure 2: Road partly masked by trees and shadows.

3.4 Correlation DSM and shadows

As shown in figure 3, heights of the DSM are often over estimated in shadow areas, leading to building false detection. As a consequence, detected shadow areas will be excluded from the change detection process, once the land cover classification will have been obtained.

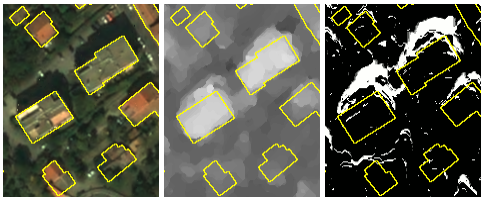


Figure 3: Heights of the DSM often over estimated in shadows. (From left to right, ortho-image, DSM, shadow mask computed from DSM)

4 APPROACH "LAND COVER CLASSIFICATION"

This approach consists in extracting land cover (and thus buildings) from satellite ortho-images and additional information using a supervised classification method. Training regions for building and road classes can here be obtained directly from the database to update. The legend of the land cover classification consists of 8 classes : "red roofed buildings", "grey roofed buildings", "white roofed buildings", "roads", "high vegetation", "low vegetation", "bare soil" and "shadows". Changes are then detected by comparing detected buildings to database objects.
The input channels for the classification task are made up by a vegetation index *"ndvi"* (calculated from red and near infrared bands of the image), the *"red"* band and the *"blue"* band.

4.1 Classification algorithms

The ortho-image is first segmented into homogeneous regions. These regions are then classified through a *MAP* classification algorithm taking into account additional knowledge as prior probabilities. The classification tool described in (Trias-Sanz, 2006) has been used for these experiments.

Model estimation from training data First, for each class, the best parameters of several statistical distributions (such as parametric laws or histograms filtered by kernel density estimation...) are computed to fit to the radiometric n-dimensional histogram of the class (with n number of used channels). Then the best model is selected thanks to Bayes Information Criterion to find a compromise between the fitting to data and the modelling complexity.

Segmentation The image must be segmented into homogeneous land cover regions. This is achieved thanks to the multi-scale segmentation method described in (Guigues et al., 2006). A pyramid of segmentations of the image is first computed. Each level of this pyramid corresponds to an alternative between detail and generalization. This pyramid is then cut at a level empirically chosen to obtain a suitable image partition. The choice of this level is a compromise between desired details and the size of regions. On one hand, in an over segmentation, some regions will be too small to have meaning and are at risk to be misclassified whereas on the other hand, in a too coarse segmentation, large regions will contain different land cover items (such as grey roofed buildings and roads here).

Classification The segmented regions are then classified according to the probability model of the radiometry of the different classes previously estimated. In the present case, a *MAP* per region classification algorithm is used. Such a method allows to take easily into account external information (see 4.2) as prior probability. With this classification method, the label $c_o(R)$ given to a region R is its most probable class according to the radiometric model previously estimated and to prior probabilities. Hence, $c_o(R)$ is the class c that maximizes the following function
$\prod_{i \ extern \ information \ source} (P_i(c(R) = c))^{a_i} \times$
$\left(\prod_{pixel \ p \in R} P_{radiometricmodel}(I(p)|c(p) = c)\right)^{\frac{1}{Card \ R}}$
with $I(p)$ standing for the radiometry vector of pixel p, $c(z)$ meaning region or pixel "z's class" and $P(c(z) = c)$ standing for the probability for pixel or region z to belong to class c. The a_i terms stand for weight parameters balancing the different prior probability sources.

4.2 Additional information

As previously said, some external knowledge will be introduced into the classification process as prior probabilities to help to prevent misclassifications when radiometric information is not sufficient.

Knowledge from the database to update It can be assumed that land cover has mostly not changed since the last time the database was updated. Such information can be taken into account as a prior probability that a pixel still belongs to the same class as in the database. Nevertheless, this information should be weighted slightly to avoid missing changes...

Ground / above ground information from nDSM Knowledge about the distinction between ground and above ground objects can be obtained from the difference nDSM between a DSM (in case of a stereoscopic acquisition of images) and a DTM. Such information helps to discriminate roads from grey roofed buildings, and bare ground from some red roofed buildings.
It is taken into account directly in the classification process as a prior probability to find the different classes according to the nDSM. Piecewise linear functions are used to model this relation (as shown by figure 4). The parameters have been selected empirically (knowing the common heights of buildings and the "precision" of the nDSM).

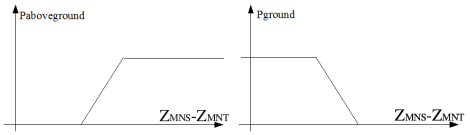


Figure 4: Probability models to belong to above-ground ("building","high vegetation") (on the left) and to ground ("road", "low vegetation","bare soil") (on the right) according to nDSM.

Knowledge about shadows Prior information about the presence of shadows can also be calculated if a DSM is available and if the capture time of images is known. Such information is useful to prevent some misclassifications between shadows and other dark objects. (A shadow mask can not be computed accurately and used directly since the correlation DSM calculated from the satellite images is quite noisy, as shown in part 3.4 and figure 3.)

Ground / above ground information from shadows Very approximate ground / above ground information for neighbouring regions of shadows can be derived from sun orientation (calculated from image capture time) and a shadow mask obtained from a previous classification.

4.3 Change detection

In this study, as it is only aimed at launching an alarm when a change concerning building theme is detected, "change detection" step is very simple : "**new**" and "**demolished building parts**" masks are computed by comparing the building mask obtained from land cover classification to the mask of the rasterized objects of the database.
An opening morphological operator is then applied to these masks in order to filter too fine objects. Other remaining small objects are eliminated if their area is lower than a given threshold.
An alarm is then launched (and checked by an operator) for each object (connected component) of these two masks. Grouping and generalizing these alarms are also possible (see part 6.2).

4.4 Practical use of this process

As previously said, the classification method is a supervised one and therefore requires training data. Training regions for building and road classes are here obtained directly from the database to update, but require pre-processing. (For other classes, a training set is captured by an operator.)

Classify buildings into sub-classes Experiments have shown that using three subclasses "red", "grey" and "white roofed buildings" instead of a single building class is necessary. However, no information concerning roofing material (and therefore roof colour) is associated to BDTopo building objects whereas such knowledge is necessary to use database objects as training data. Nevertheless this information can be obtained from image data. Every pixel p of the ortho-image belonging to a building object of the database is thus classified as "red", "grey" or "white" using the following scheme :

if $\frac{p.red}{p.blue} > threshold_{redVSgrey\_white}$ **then** p is "red"
else if $p.red > threshold_{greyVSwhite}$ **then** p is "white"
else p is "grey" **endif**.

Clean training data As already said in 3.4, roads and building objects often appear partly masked by higher objects (trees) or shadows on the image. Besides, database is quite generalized and therefore a database object can include few parts of neighbouring objects. Therefore, these training regions are quite "dirty". In order to cope with this, the classification is performed in a two pass scheme :

1. Ortho-image is classified according to a model estimated from the rough database without any special care.

2. Training data is cleaned using vegetation and shadow areas masks obtained from this first classification. A new model is then estimated and a new classification is performed.

5 APPROACH "INTERDATE DSM COMPARISON"

This approach consists in comparing a DSM computed from stereoscopic satellite images (at date $t2$) to a DSM describing the landscape at time ($t1$) when the database was last updated. That is to say, their difference $dDSM$ is studied, since new building parts correspond to zones where $Z_{DSM_{t2}} > Z_{DSM_{t1}}$, whereas demolished ones correspond to $Z_{DSM_{t2}} < Z_{DSM_{t1}}$. (Besides this measure *dDSM* could also be a confidence measure associated to each detected change.) Nevertheless, other information (such as a vegetation mask obtained at date $t2$ from previous classification of new images) must be taken into account to prevent false alarms. The tested algorithm is described below :

for each pixel p **do**
  $dDSM(p) = Z_{DSM_{t2}}(p) - Z_{DSM_{t1}}(p)$
  if $dDSM(p) < th_{demolished}$ **and** $p \in$ { building object of the database } **then**
    $p \in$ {demolished building parts}
  else if $dDSM(p) > th_{new}$ **and** p not classed among shadows and vegetation **then**
    if $p \in$ { building object of the database } **then**
      No "2D" change
    else
      $p \in$ {new building parts}
    endif
  endif
endfor

Perform morphological opening of "demolished building parts" and "new building parts" masks and eliminate too small remaining elements (connected components) ($area < threshold$) of these masks.

In the present study, as no old DSM was available, a "pseudo-DSM" has been generated from a DTM and from height attributes of database building objects, in order to describe landscape at the time the database was last updated.

6 USE ALARMS

6.1 Merge alarms

Results obtained by both approaches are merged in a very simple way. First, intersection and union masks of alarms launched by both approaches are computed. Secondly, too small elements of the intersection mask are deleted. Lastly, objects of the union mask are deleted if they don't contain an object of the intersection mask.
This simple method is a way to reduce the number of alarms to check. It improves the correctness rate, but **requires a high completeness** for alarms launched by the both approaches. It could be extended to a more general scheme giving confidence weights to different alarms sources (and even associating confidence scores to each alarm). This method is a way to obtain a mask of alarms but also a confidence score related to the recurrence of alarms and to belief given to the different sources of alarms.
Besides, as some existing methods work at building level (such as (Rottensteiner, 2007) or (Champion et al., 2010)), alarms could be associated to a building (new, demolished or modified) and a

confidence score computed for each building (using the scheme presented above), making it possible to sort alarms from the most to the least plausible.

6.2 Group/generalize alarms

Group/generalize alarms is an other way to get less alarms to check. Alarms located in a same neighbourhood are grouped into a single object. These patterns are then submitted to an operator to be checked. This alarm generalization (artificially) improves correctness rate without damaging completeness rate, especially in cases such as the one illustrated by figure 5, where a smaller amount of more global alarms can then be submitted to operators.

Figure 5: Group neighbouring alarms is a way to obtain less alarms to check (especially in cases like this one...)

7 TESTS AND RESULTS

Results are presented and commented for both datasets. Completeness and correctness rates are presented in tables 1 and 2. Concerning computing time, processing the whole chain takes almost 1 h/km^2.

7.1 Toulouse

As available database was assumed to be up-to-date, simulated changes consisting in adding or deleting buildings in the database were performed in order to be able to easily evaluate change detection results. But it appeared that there were true changes and bad plotted or misregistered buildings. As a consequence, the correctness rate has been computed as the ratio of "manually" validated alarms compared to the number of launched alarms, whereas the completeness rate has been calculated as the ratio of detected simulated changes among the simulated changes.
Furthermore, as database objects are quite generalized, it is sometimes not easy to decide whether an alarm corresponds to a true change (or a bad modelled object) and must be validated.

Approach "land cover classification" For first tests, a monoscopic acquisition case was assumed. Therefore, no DSM (and therefore no information concerning ground or overground) was available. White roofed buildings have been well detected, while grey roofed buildings and roads have very often been misclassified. Confusion has also occurred between red roofed buildings and bare soil without this knowledge.
A stereoscopic acquisition was assumed for other tests. It has then been showed that the results are greatly improved by the use of nDSM derived knowledge concerning ground/overground. Some results are presented in tables 1 and 2. (They have been obtained taking into account radiometric model, nDSM knowledge (weighted with 0.75 and 0.25 respectively).
Approach "DSM comparison" Obtained results are good.
Merge results Merge alarms hugely improves the correctness rate for these areas, without damaging the completeness rate, thanks to the high completeness of both approaches here.

Conclusion Interesting results have been obtained, but it must be kept in mind that Toulouse test areas remain quite simple ones since there are few classes (mostly red roofs, no water areas or dark slate roofs for instance) and relief is not very strong... Besides, used databases were high quality ones.

7.2 La Réunion

BDTopo database has recently been updated by IGN operators in this area. Therefore, the old database and the modifications plotted by operators were available. Nevertheless, as for Toulouse test zones, some few errors were also detected on a small area, showing that change detection remains a difficult task and that may be not perfect even when processed by operators, as shown in figure 6.

Obtained results (presented in tables 1 and 2) on this area are not as good as the ones obtained for Toulouse test zones, but this can be explained by several reasons. First, the correlation DSM is quite noisy : indeed it has been generated from images not captured at the same date (but at one week apart with a low $\frac{B}{H}$ ratio 0.39). Furthermore, it has been computed from already pansharpened colour images, and not directly from original (high resolution) panchromatic images. Besides, roofing materials are often metallic ones leading to specular reflectance and saturated pixels. Secondly, many buildings are very low (only one level) and are therefore difficult to detect as above ground objects in the nDSM : they are lost among noise of the nDSM, whereas this ground/above-ground knowledge was a key factor for a good building detection on Toulouse test zones. Furthermore, relief is also stronger.
Last, roofing materials are different and more various than in Toulouse. Furthermore, some buildings appear almost as dark as shadows or roads on satellite images (but not on available older aerial ortho-images on which they are easier to distinguish. It must here be said that these GeoEye-1 images have also a quite different radiometry from Pleiades simulated ones). These buildings are thus often missed leading to many false positives for demolished objects and to true negatives for new objects.
All these phenomenons have effects on both approaches "land cover classification" and "DSM comparison", requiring to modify some parameters (thresholds for colour classification of buildings, and parameters of ground/above ground according to nDSM). The completeness rate is not so low, even though many dark roofed buildings are missed. The correctness is not so bad, but it must be kept in mind that the number of launched alarms is important (up to 687). Concerning demolished buildings, the correctness rate is very weak (almost 2%) while too many (mostly false) alarms (353) have been launched for an operational use.

Conclusion La Réunion is a difficult test compared to Toulouse (both concerning available data and landscape), and obtained results are not so good. Nevertheless, it must be kept in mind that this landscape is quite different (low buildings) from the main part of the French territory.

8 CONCLUSION

Automatic change detection from satellite images remains a difficult task. Nevertheless, although they are still not suitable for an operational use, obtained results remain encouraging, considering input data. A very good completeness rate is indeed reached for Toulouse test areas (almost 100%), while still "interesting" one (almost 75%) is obtained for La Réunion (since it revealed that some true changes have been missed by operators on a very small area).
One important limit for an operational use is the bad correctness rate ($< 30\%$): too many false alarms are launched. Nevertheless, this can be improved by generalizing alarms, submitting to validation patterns of alarms instead of isolated alarms (such as on figure 5). Thus, the correctness rate increases while completeness rate remains similar. (Using a 10 meter radius neighbourhood to group patterns of alarms, the correctness rate increased from 48%

Table 1: Completeness rates for approaches "classif" (land cover classification), "dDSM" (interdate DSM comparison), and "merged" (merged alarms of both methods). Number of simulated (new/demolished) objects is given in the second column.

area	nb simul	classif	dDSM	merged
New buildings				
Highway	30	97%	93%	93%
CHU	10	100%	100%	100%
Réunion	251	80%	74%	71%
Demolished buildings				
Highway	5	100%	100%	100%
CHU	7	100%	100%	100%

Table 2: Correctness rates for approaches "classif", "dDSM" and "merged" alarms.

area	classif	dDSM	merged
New buildings			
Highway	40% (257)	48% (201)	56% (128)
CHU	30% (193)	19% (242)	46% (98)
Réunion	48% (509)	33% (687)	65% (321)
Demolished buildings			
Highway	14% (57)	73% (11)	100% (5)
CHU	15% (59)	36% (20)	70% (10)

to 52% in La Réunion test area.)

An other possible solution consists in merging results obtained by this approach with results obtained from other methods such as (Champion et al., 2010) for instance : confidence indices can then be calculated for each alarm and only the most relevant ones would be submitted to validation. Other sources such as crowdsourcing could be used too.

Completeness seems to be a very important point too since it is important to be exhaustive, to be sure not to miss changes. Nevertheless, this criterion could become less important in case of a more important image acquisition frequency. Intermediate satellite images could now be acquired between two consecutive (4 years apart) aerial images acquisition campaigns. Therefore, there would be less time to treat them and some (semi-)automatic change detection tools would become necessary to cope with such amount of data. As a consequence, completeness would become less important than nowadays, since even if change detection was not perfect but sufficiently good for these intermediate images, some few changes would be missed by (semi-)automatic tools whereas they would be detected afterwards when a complete investigation would be performed on aerial images by an operator two years later. Investigation by operators remains indeed necessary, since this is the only way for the moment to detect changes for some other themes than buildings and roads. However, presented automatic tools could help change detection or database verification, even though they still need "manual" investigation.

Interesting results have been obtained but more experiments remain necessary, for instance to know the robustness to parameters and to different kinds of landscape, e.g. with additional themes such as water or slate (dark blue) roofed buildings. An other important point has been shown : even the first approach remains dependent on "sufficiently good" 3D information (that is to say on the correlation DSM) to be able to discriminate between ground and above ground classes with quite similar radiometry and thus requires stereoscopic acquisition.

ACKNOWLEDGEMENTS

This work was partly financed by CNES. The authors would also like to thank M. Durupt and X. Maranzana for their help especially during data preparation.

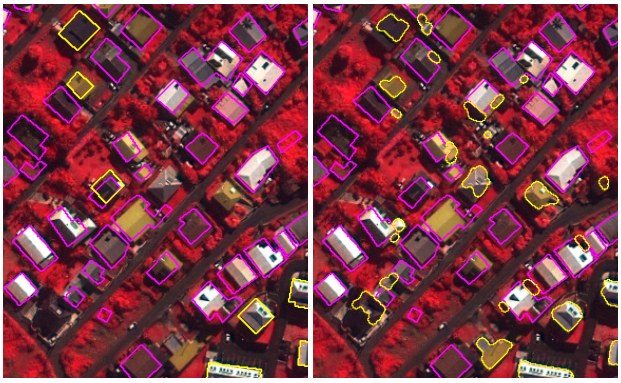


Figure 6: Change detection example in La Réunion. Old database is drawn in magenta. On the left, modifications (in green) plotted by IGN operators. On the right, launched alarms (in green) for detected "new building parts" by the "land cover classification" approach.

REFERENCES

Bouziani, M., Goïta, K. and He, D.-C., 2010. Automatic change detection of buildings in urban environment from very high spatial resolution images using existing geodatabase and prior knowledge. ISPRS Journal of Photogrammetry and Remote Sensing 65, pp. 143–153.

Champion, N., Boldo, D., Pierrot-Deseilligny, M. and Stamon, G., 2010. 2D building change detection from high resolution satellite imagery : A two-step hierarchical method based on 3D invariant primitives. Pattern Recognition Letters 31, pp. 1138–1147.

CNES, 2011. Pleiades. http://smsc.cnes.fr/PLEIADES/index.htm.

GeoEye, 2011. About GeoEye-1. http://launch.geoeye.com/LaunchSite/about/.

Guigues, L., Cocquerez, J.-P. and Le Men, H., 2006. Scale sets image analysis. International Journal of Computer Vision 68(3), pp. 289–317.

Matikainen, L., Hyyppä, J., Ahokas, E., Markelin, L. and Kaartinen, H., 2010. Automatic detection of buildings and changes in buildings for updating of maps. Remote Sensing 2(5), pp. 1217–1248.

Olsen, V. and Knudsen, J., 2005. Automated change detection for validation and update of geodata. In: Proceedings of 6th Geomatic Week, Barcelone, Spain.

Pierrot-Deseilligny, M. and Paparoditis, N., 2006. A multiresolution and optimization-based image matching approach: An application to surface reconstruction from SPOT5-HRS stereo imagery. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (IAPRS), Vol. 36 (1/W41), Ankara, Turkey.

Poulain, V., Inglada, J., Spigai, M., Tourneret, J.-Y. and Marthon, P., 2009. Fusion of high resolution optical and SAR images with vector data bases for change detection. In: IEEE International Geoscience and Remote Sensing Symposium, Cape Town, South Africa.

Radke, R., Andra, S., Al-Kofahi, O. and Roysam, B., 2005. Image change detection algorithms: A systematic survey. IEEE Transactions on image processing 14(3), pp. 294–307.

Rottensteiner, F., 2007. Building change detection from digital surface models and multi-spectral images. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (IAPRS), Vol. 36 (3/W49B), Munich, Germany.

Trias-Sanz, R., 2006. Semi-automatic high-resolution rural land cover classification. PhD thesis, Université Paris 5, Paris, France.

DETECTION OF WINDOWS IN IR BUILDING TEXTURES USING MASKED CORRELATION

D. Iwaszczuk, L. Hoegner, U. Stilla

Photogrammetry & Remote Sensing, Technische Universitaet Muenchen (TUM), Arcisstr. 21, 80333 Munich, Germany – (iwaszczuk, hoegner, stilla)@bv.tum.de

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: infrared, image sequences, texture mapping, structure detection

ABSTRACT:

Infrared (IR) images depict thermal radiation of physical objects. Imaging the building hull with an IR camera allows thermal inspections. Mapping these images as textures on 3D building models, 3D geo-referencing of each pixel can be carried out. This is helpful for large area inspections. In IR images glass reflects the surrounding and shows false results for the temperature measurements. Consequently, the windows should be detected in IR images and excluded for the inspection. In this paper, an algorithm for window detection in textures extracted from terrestrial IR images is proposed. First, a local dynamic threshold is used to extract candidates for windows in the textures. Assuming a regular grid of windows masked correlation is used to find the position of windows. Finally, gaps in the window grid are replaced by hypothetical windows. Applying the method for a test dataset, 79% completeness and 80% correctness was achieved.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

FAST MARCHING FOR ROBUST SURFACE SEGMENTATION

F. Schindler, W. Förstner

Department of Photogrammetry, University of Bonn, Nussallee 15, 53115 Bonn, Germany –
falko.schindler@uni-bonn.de, wf@ipb.uni-bonn.de

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: segmentation, point clouds, laser range data

ABSTRACT:

We propose a surface segmentation method based on Fast Marching Farthest Point Sampling designed for noisy, visually reconstructed point clouds or laser range data. Adjusting the distance metric between neighboring vertices we obtain robust, edge-preserving segmentations based on local curvature. We formulate a cost function given a segmentation in terms of a description length to be minimized. An incremental-decremental segmentation procedure approximates a global optimum of the cost function and prevents from under- as well as strong over-segmentation.We demonstrate the proposed method on various synthetic and real-world data sets.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

A PERFORMANCE STUDY ON DIFFERENT STEREO MATCHING COSTS USING AIRBORNE IMAGE SEQUENCES AND SATELLITE IMAGES

K. Zhu[1], P. d'Angelo[2], M. Butenuth[1]

[1] Technische Universitaet Muenchen (TUM), Remote Sensing Technology – (ke.zhu, matthias.butenuth)@bv.tum.de
[2] German Aerospace Center (DLR), Remote Sensing Technology Institute – pablo.angelo@dlr.de

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: dense stereo matching, cost function, performance, observation constraints

ABSTRACT:

Most recent stereo algorithms are designed to perform well on close range stereo datasets with relatively small baselines and good radiometric conditions. In this paper, different matching costs on the Semi-Global Matching algorithm are evaluated and compared using aerial image sequences and satellite images with ground truth. The influence of various cost functions on the stereo matching performance using datasets with different baseline lengths and natural radiometric changes is evaluated. A novel matching cost merging Mutual Information and Census is introduced and shows the highest robustness and accuracy. Our study indicates that using an adaptively weighted combination of Mutual Information and Census as matching cost can improve the peformance of stereo matching for airborne image sequences and satellite images.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

FUSION OF DIGITAL ELEVATION MODELS USING SPARSE REPRESENTATIONS

H. Papasaika[1], E. Kokiopoulou[2], E. Baltsavias[1], K. Schindler[1], D. Kressner[3]

[1] Institute of Geodesy and Photogrammetry, ETH Zurich, Switzerland
[2] Seminar for Applied Mathematics, ETH Zurich, Switzerland
[3] Mathematics Institute of Computational Science and Engineering, EPFL, Switzerland

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: DEMs, fusion, quality evaluation, learning

ABSTRACT:

Nowadays, different sensors and processing techniques provide Digital Elevation Models (DEMs) for the same site, which differ significantly with regard to their geometric characteristics and accuracy. Each DEM contains intrinsic errors due to the primary data acquisition technology, the processing chain, and the characteristics of the terrain. DEM fusion aims at overcoming the limitations of different DEMs by merging them in an intelligent way. In this paper we present a generic algorithmic approach for fusing two arbitrary DEMs, using the framework of sparse representations. We conduct extensive experiments with real DEMs from different earth observation satellites to validate the proposed approach. Our evaluation shows that, together with adequately chosen fusion weights, the proposed algorithm yields consistently better DEMs.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

CHANGE DETECTION IN URBAN AREAS BY DIRECT COMPARISON OF MULTI-VIEW AND MULTI-TEMPORAL ALS DATA

M. Hebel[1], M. Arens[1], U. Stilla[2]

[1] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation
IOSB, Gutleuthausstr. 1, 76275 Ettlingen, Germany –
(marcus.hebel, michael.arens)@iosb.fraunhofer.de
[2] Department of Photogrammetry and Remote Sensing, Technische Universitaet Muenchen (TUM), Germany –
stilla@tum.de

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: airborne laser scanning, LiDAR, change detection, multi-temporal data analysis, urban areas

ABSTRACT:

Change detection in urban areas requires the comparison of multi-temporal remote sensing data. ALS (airborne laser scanning) is one of the established techniques to deliver these data. A novelty of our approach is the consideration of multiple views that are acquired with an oblique forward-looking laser scanner. In addition to advantages in terms of data coverage, this configuration is ideally suited to support helicopter pilots during their mission, e.g., with an obstacle warning system, terrain-referenced navigation, or online change detection. In this paper, we present a framework for direct comparison of current ALS data to given reference data of an urban area. Our approach extends the concept of occupancy grids known from robot mapping, and the proposed change detection method is based on the Dempster-Shafer theory. Results are shown for an urban test site at which multi-view ALS data were acquired at an interval of one year.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

TOWARDS AIRBORNE SINGLE PASS DECIMETER RESOLUTION SAR INTERFEROMETRY OVER URBAN AREAS

M. Schmitt[1], C. Magnard[2], T. Brehm[3], U. Stilla[1]

[1] Photogrammetry & Remote Sensing, Technische Universitaet Muenchen (TUM), Munich, Germany – michael.schmitt@bv.tum.de, stilla@tum.de
[2] Remote Sensing Laboratories, University of Zurich, Zurich, Switzerland – christophe.magnard@geo.uzh.ch
[3] Fraunhofer Institute for High-Frequency Physics and Radar Techniques, Wachtberg, Germany – thorsten.brehm@fhr.fraunhofer.de

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: InSAR, urban areas, very high resolution, DSM

ABSTRACT:

Airborne cross-track Synthetic Aperture Radar interferometers have the capability of deriving three-dimensional topographic information with just a single pass over the area of interest. In order to get a highly accurate height estimation, either a large interferometric baseline or a high radar frequency has to be used. The utilization of a millimeter wave SAR allows precise height estimation even for short baselines. Combined with a spatial resolution in the decimeter range, this enables the mapping of urban areas from airborne platforms. The side-looking SAR imaging geometry, however, leads to disturbing effects like layover and shadowing, which is even intensified by the shallow looking angle caused by the relatively low altitudes of airborne SAR systems. To solve this deficiency, enhanced InSAR processing strategies relying on multi-aspect and multi-baseline data, respectively, are shown to be necessary.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

REGIONWISE CLASSIFICATION OF BUILDING FACADE IMAGES

M.Y. Yang, W. Förstner

Department of Photogrammetry, University of Bonn, Nussallee 15, 53115 Bonn, Germany – michaelyangying@uni-bonn.de, wf@ipb.uni-bonn.de

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: classification, conditional random field, random decision forest, segmentation, facade image

ABSTRACT:

In recent years, the classification task of building facade images receives a great deal of attention in the photogrammetry community. In this paper, we present an approach for regionwise classification using an efficient randomized decision forest classifier and local features. A conditional random field is then introduced to enforce spatial consistency between neighboring regions. Experimental results are provided to illustrate the performance of the proposed methods using image from eTRIMS database, where our focus is the object classes building, car, door, pavement, road, sky, vegetation, and window.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

SUPERVISED CLASSIFICATION OF MULTIPLE VIEW IMAGES IN OBJECT SPACE FOR SEISMIC DAMAGE ASSESSMENT

M. Gerke

University of Twente, Faculty of Geo-Information Science and Earth Observation –
ITC, Department of Earth Observation Science, Hengelosestraat 99, P.O. Box 217, 7500AE Enschede, Netherlands

Working Groups I/2, III/1, III/4, III/5

KEYWORDS: AdaBoost, classification, feature, fusion, learning, performance, point cloud, random trees

ABSTRACT:

Classification of remote sensing image and range data is normally done in 2D space, because anyhow most sensors capture the surface of the earth from a close-to vertical direction and thus vertical structures, e.g. at building facades are not visible anyways. However, when the objects of interest are photographed from off-nadir directions, like in oblique airborne images, the question on how to efficiently classify those scenes arises. In this paper a study on classification in 3D object space is presented: image features from individual oblique airborne images, and 3D geometric features derived from matching in those images are projected onto voxels. Those are segmented and classified. The study area is Port-Au-Prince (Haiti), where images have been acquired after the earthquakes in January 2010. Results show that through the combination of image evidence as realized by the projection into object space the classification becomes more accurate compared to single image classification.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

CONDITIONAL RANDOM FIELDS FOR URBAN SCENE CLASSIFICATION WITH FULL WAVEFORM LIDAR DATA

J. Niemeyer[1], J.D. Wegner[1], C. Mallet[2], F. Rottensteiner[1], U. Soergel[1]

[1] Institute of Photogrammetry and GeoInformation, Leibniz Universitaet Hannover, Nienburger Str. 1, 30167 Hannover, Germany – (niemeyer, wegner, rottensteiner, soergel)@ipi.uni-hannover.de
[2] Laboratoire MATIS, Institut Geographique National, Universite Paris Est, 73 avenue de Paris, 94165 Saint-Mandé, France – clement.mallet@ign.fr

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: conditional random fields, 3D point cloud, full waveform LiDAR, urban, classification

ABSTRACT:

We propose a context-based classification method for point clouds acquired by full waveform airborne laser scanners. As these devices provide a higher point density and additional information like echo width or type of return, an accurate distinction of several object classes is possible. However, especially in dense urban areas correct labelling is a challenging task. Therefore, we incorporate context knowledge by using Conditional Random Fields. Typical object structures are learned in a training step and improve the results of the point-based classification process. We validate our approach with two real-world datasets and by a comparison to Support Vector Machines and Markov Random Fields.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

OBJECT-BASED FOREST CHANGE DETECTION USING HIGH RESOLUTION SATELLITE IMAGES

Nesrine Chehata [1], Camille Orny [1,2], Samia Boukir [1] and Dominique Guyon [2]

1) G&E Laboratory, ENSEGID, Bordeaux University,
1 Allée F. Daguin, 33607 Pessac Cedex, France
2) EPHYSE Laboratory, INRA,
33140 Villenave d'Ornon, France
nesrine.chehata@egid.u-bordeaux3.fr

Working Groups III/4,III/5,IIV/4,IIV/5

KEY WORDS: multitemporal classification, segmentation, feature selection, change detection, forest damage

ABSTRACT:

An object-based approach for forest disaster change detection using High Resolution (HR) satellite images is proposed. An automatic feature selection process is used to optimize image segmentation via an original calibration-like procedure. A multitemporal classification then enables the separation of wind-fall from intact areas based on a new descriptor that depends on the level of fragmentation of the detected regions. The mean shift algorithm was used in both the segmentation and the classification processes. The method was tested on a high resolution Formosat-2 multispectral satellite image pair acquired before and after the Klaus storm. The obtained results are encouraging and the contribution of high resolution images for forest disaster mapping is discussed.

1 INTRODUCTION

In a climate changing context, wind storms have become more and more frequent. Wind-fall damages have to be quickly mapped to prevent fire risks, call for financial compensation and to update the national forest inventory. While ground investigations are complex due to fallen trees, remote-sensing techniques enable fast monitoring of large and unreachable areas. Their use have widely spread through the world for disaster change detection, especially with the growing spatial and temporal resolutions of new satellites. The objective of the present study is to provide a binary mapping discriminating damaged and intact areas using HR satellite imagery consisting of a pair of multispectral Formosat-2 images of 8 m resolution. The images were taken before and after the Klaus Storm that happened the 24^{th} January 2009 in the South West of France.

In literature, the former works in forestry produced low scale maps, near to the hectare and studied essentially the clear-cuts. Few works have dealt with smaller structural changes such as wind-fall damages. Recently, object-based classifications were used for change detection in forestry (Desclee et al., 2006, Conchedda et al., 2008). These methods are based on a segmentation process that combines spatial and spectral information to group pixels into homogeneous regions before their classification using new object descriptors. The latter can be geometrical and textural (Fraser et al., 2005) or temporal (Desclee et al., 2006).

Besides, change detection can be either based on a comparison of before and after storm classifications (post-classification), or directly processed on multitemporal images (joint-classification). Post-classification approaches are robust to radiometric differences between images, and provide an accurate "from-to" change information (Im and Jensen, 2005) but suffer from errors propagation. Joint-classification approaches provide more information to classify small changes. They can either rely on machine learning algorithms using a training set (Im and Jensen, 2005) or use thresholding which involves a parametric statistical test (Desclee et al., 2006) or an expert knowledge (Fraser et al., 2005).

In this study, the proposed method is an object-based, multitemporal classification that maps storm damages at a fine spatial scale. We propose a nearly automatic method requiring very limited data for rapid mapping at a regional scale. An unsupervised multitemporal classification was preferred to a post-classification scheme since, in our context, changes are subtle and hard to detect in a single after-storm image. The algorithm is based on the mean shift segmentation that will be detailed in section 2. An automatic feature selection process is used to optimize image segmentation via an original calibration-like procedure and will be presented in section 3. Section 4 presents the binary multitemporal classification which is based on the mean shift algorithm and uses a new descriptor that depends on the level of fragmentation of the detected regions. Experimental results are shown and discussed in section 5 and finally conclusions are drawn.

2 MEAN SHIFT SEGMENTATION

The Mean Shift (MS) algorithm is a non-parametric feature-space analysis technique. (Fukunaga and Hostetler, 1975, Comaniciu and Meer, 2002) showed excellent results in clustering and object delineation in color images. It is based on a density mode searching and clustering technique. The feature space is considered as the empirical probability density function (p.d.f.) of the input features. The algorithm proposes a filtering step that associates each pixel in the image with the closest local mode in the density distribution of the feature space. The MS procedure actually locates theses modes without estimating the global density. The segmentation into a piecewise constant structure requires one more step, the fusion of the regions associated with nearby modes. The implementation of (Comaniciu and Meer, 2002) searches for local modes in the joint feature and spatial domain of $n + 2$ dimensions, where n is the number of considered features. An iterative procedure of mode seeking consists in shifting the n+2 dimensional window to a local mode. The search window involves two user-defined inputs that can be deduced from desired object sizes or physical properties. A radiometric range (h_r) corresponds to the unique spectral radius in the n-dimensions search window and a spatial bandwidth (h_s) corresponds to the spatial radius of the window.

Figure 1 shows the impact of both parameters on segmentation results.

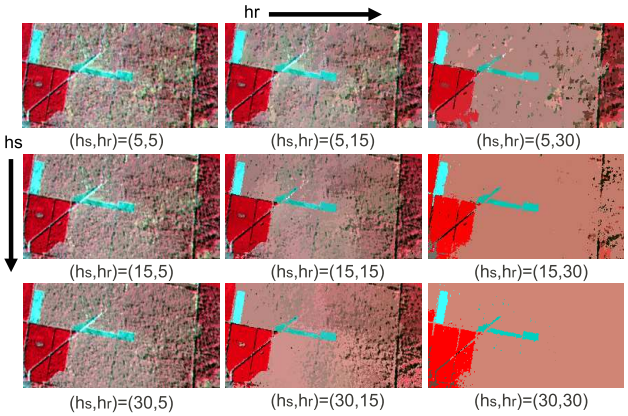


Figure 1: Mean shift segmentations of a 4-channel image (B,G,R,NIR) using different parameters (h_s, h_r). When h_r increases, only highly contrasted and homogeneous regions remain: intact young stand on the left, blue farming area, medium aged stand on the right. If h_s increases, only larger regions remain : shadow areas and standing tree groups disappear within the central, strongly damaged area.

Unlike hierarchical methods used in (Desclee et al., 2006, Conchedda et al., 2008), the Mean Shift does not use any heterogeneity measure which is more appropriate in forestry where the objects of interest, i.e. tree stands, include heterogeneous pixels such as vegetation, ground and shadows. To our knowledge, the MS algorithm has not been used for forestry mapping. In our method, the Mean Shift procedure is used in both segmentation and classification steps using the joint spatio-spectral domain and solely the spectral domain respectively (*cf.* Section 4).

3 AUTOMATIC FEATURE SELECTION FOR SEGMENTATION

3.1 Feature selection using test frames

In literature, the segmentation is usually processed on all available before and after image bands. In this study, in order to determine the most relevant features for segmentation, input feature selection was carried out through an original generic calibration-like procedure using a test frame. Moreover, this process aims to automatically optimize the segmentation parameters. The test frames are constructed with n small image samples corresponding to n classes, yielding as much test frames as input features, thus resulting in a *multi-band test frame*, at the core of our feature selection method. The used data is not included in the validation dataset. The test frames are then segmented by the MS algorithm using either one input feature (single test frame) or multiple normalized features (multi-band test frame) while testing various parameterizations. The feature Segmentation Performance (SP) is defined as:

$$SP = \frac{1}{n_c} \sum_{i=1}^{c} \frac{\max_j [A(R_i) \cap A(R_{S_j})]}{A(R_i)} \ , \ R_i \cap R_{S_j} \neq \emptyset \quad (1)$$

where A is the area, n_c the actual number of classes in the test frame, R_i the actual regions and R_{S_j} the segmented regions. The segmentation performance depends on the regions overlap percentage. Figure 2 illustrates an example of a test frame with three regions and the SP computation.

The highest SP value leads to the best feature or set of features.

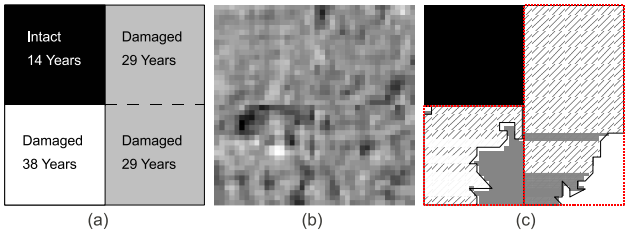


Figure 2: (a) Actual regions, (b) Blue band test frame, (c) Segmented regions using $(h_s, h_r) = (5, 10)$, maximal overlaps per region are hatched. Segmentation performance SP=86%.

3.2 Proposed input features

Forest features can be grouped into three categories: spectral, textural and temporal features. Spectral features include raw or corrected spectral bands and derived bands such as the NDVI (Normalized Difference Enhanced Vegetation Index) (Conchedda et al., 2008), the PCI (Principal Components Inversion) (Inglada, 2001), etc. In this study, the four Formosat-2 spectral bands (blue, green, red and infrared) were used as well as two vegetation indexes, the Normalized Difference and Soil Adjusted Vegetation Indexes (NDVI and SAVI). First order statistics such as mean or variance of reflectance are also used. As for textural features, the more common ones are Haralick features (Haralick et al., 1973) as used in (Ruiz et al., 2004, Kayitakire et al., 2006, St-Louis et al., 2006, Trias-Sanz et al., 2008, Tuominen and Pekkarinen, 2005). Finally, three common temporal features were computed: mean correlation, difference and ratio between both images. More temporal features can be found in literature such as the pixel-wise Magnitude of the Temporal Change Vector (Fraser et al., 2005), the multi-temporal PCI (Inglada, 2001), the Neighborhood Correlation Index (Im and Jensen, 2005) and will be investigated in a future work. Temporal and textural features were then processed for each spectral feature.

4 UNSUPERVISED OBJECT-BASED MULTITEMPORAL CLASSIFICATION

The global unsupervised object-based multitemporal change detection scheme is depicted in Figure 3. The before and after-storm images are segmented independently using the MS algorithm and the feature selection process as explained in Section 3. The before-storm segments correspond to homogeneous structural regions, i.e belonging to the same age class. The after-storm segments reflect the change degree.

4.1 Mean Shift spectral classification

The after-storm segmented regions are represented by object mean temporal descriptors. The automatic feature selection process (*cf.* Section 3.1) provides the input features that optimize the MS spectral classifier. The segmented regions are then clustered automatically into change classes using this optimized mean shift spectral classifier. Unlike the MS segmentation, this modified version is independent of pixel positions and involves solely the spectral domain which allows to cluster similarly damaged regions that are spatially distant into the same change class. The MS classifier has a single parameter, h_r, which is easy to specify. The MS spectral classification leads automatically to many change clusters. No reference data were available to validate the obtained change degrees. Consequently, our objective was limited to the production of a binary change map even if the MS spectral classification provides multiple change classes.
In order to group the change classes into intact and damaged classes, the clusters were characterized by an innovative temporal feature : the fragmentation rate.

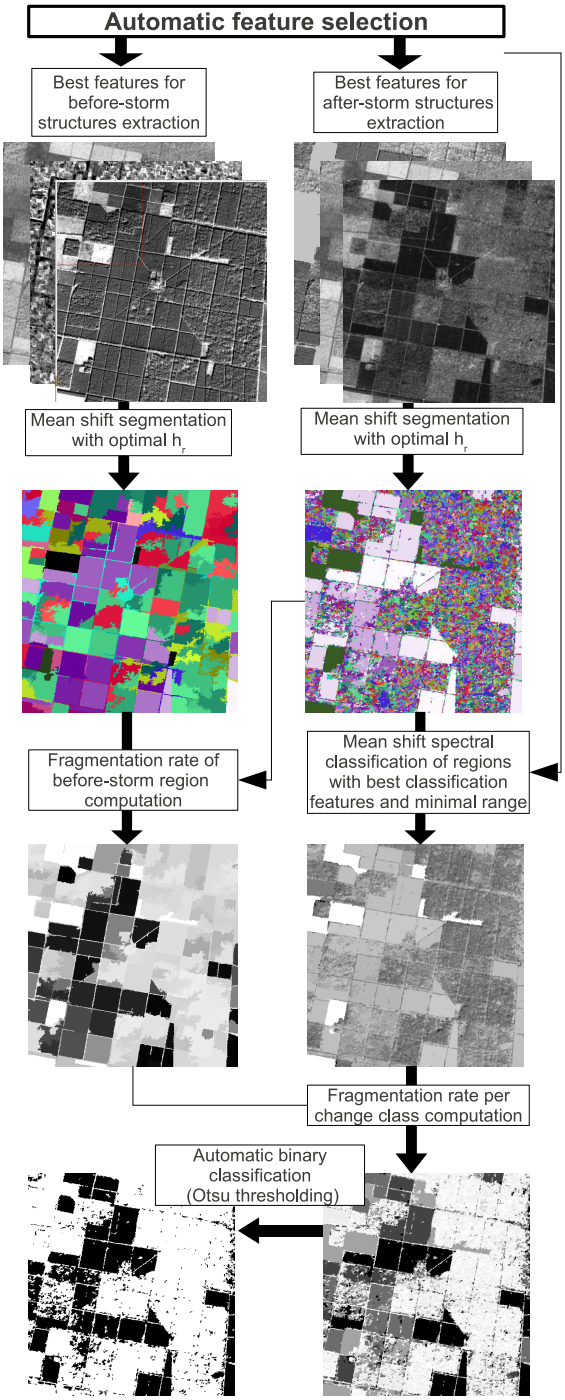


Figure 3: Multitemporal object-based change detection scheme.

4.2 Fragmentation rate

Damaged areas are heterogeneous and therefore appear over-segmented in the after-storm image. Conversely, intact areas correspond to larger regions and have similar delimitations in both images. The Fragmentation Rate (FR), characterizes the before-storm regions and reflects their over-segmentation in the after-storm image. It is computed by a comparison between before and corresponding after-storm regions as follows:

$$FR(R) = 1 - \frac{\max_j [A(R) \cap A(R_j^a)]}{A(R)} \ , \ R \cap R_j^a \neq \emptyset \qquad (2)$$

where A is the area, R is the before-storm region and R_j^a the after-storm regions that are included (partly or entirely) in the before-storm one.

The average *FR* is then computed for the change clusters after-storm. The change class fragmentation rate is then defined as :

$$FR(R_C) = \frac{1}{N} \sum_{p \in R_C, p \in R_i} FR(R_i) \qquad (3)$$

where R_C is the change cluster, p and N are respectively the pixels and the number of pixels of the change class C, R_i are the before-storm regions. The clusters are finally divided into two intact and damaged classes, based on their fragmentation rate and using the unsupervised Otsu thresholding (Otsu, 1979) which minimizes intra-class variances.

5 RESULTS AND DISCUSSION

5.1 Study area and data set

The Nezer forest is located on the French Atlantic coast. It is made up of rectangular stands of pine trees that have the same age and height. Available images of the area are a set of ortho-rectified, geo-referenced multispectral Formosat-2 images before and after the Klaus storm, acquired on 22/12/08 and 04/02/09, respectively (*cf.* Figure 4). The images have a 8m spatial resolution and four spectral bands (Blue, Green, Red and Near Infrared). Ancillary data include tree stand delimitations and ages GIS layer and 100 reference areas identified on orthophotos of 15 cm resolution dating back to 26/02/09.

(a) Before-storm image (b) After-storm image

Figure 4: Formosat-2 multispectral images acquired before and after the Klaus storm

5.2 Feature Selection for segmentation and classification

Table 1 outlines the used input features, separated into three groups: spectral, textural and temporal. A total of 84 features are used: 6 spectral features, 10 textural and 3 temporal features all computed on the six spectral features respectively. For the textural features, the neighborhood radius and the directional vector offsets were both experimentally set to 1 pixel, the displacement vector being horizontal.

These features were used in the feature selection process, individually or combined in a multi-feature test frame after their normalization. A test frame composed of four regions was used. The spatial radius h_s of the MS segmentation was set using a prior thematic knowledge on the desired objects size. A 3-pixel area

Spectral	Textural	Temporal
Blue	Mean	Difference
Green	Variance	Ratio
Red	8 Haralick features:	Mean correlation
PIR	mean, contrast,entropy	
NDVI	Angular Second Moment	
SAVI	Inverse Difference Moment	
	Sum average, Sum entropy	
	Sum variance	

Table 1: Input features

can reasonably be considered as a regular forest pattern, so h_s was set to 3. To set the radiometric range h_r, all features were individually rescaled between 0 and 255 and spectral ranges from 2 to 60 were experimented via our automated Mean-Shift parameter optimization procedure.

In the MS spectral classification, the smallest spectral range h_r=2 was used in order to detect subtle changes. Besides, unlike the MS segmentation, this spectral-based clustering stage involves only object temporal features. They derived from averaging the temporal features over the segmented after-storm regions. Tables 2 and 3 present the best features and spectral ranges for image segmentation and classification respectively.

After-storm segmentation			Before-storm segmentation		
Feature	h_r	SP (%)	Feature	h_r	SP (%)
Red Ratio	17	87.2	Red	2	80.7
Red	3	78.1	NDVI	3	77.3
Green Ratio	16	75.5	Green	5	68.7

Table 2: Optimal features and spectral ranges h_r for the after-storm and the before-storm segmentations using a test frame of 4 samples. SP is the segmentation performance (*cf.* Eq. 1).

Feature	SP (%)
Green Difference	92.6
NIR Difference	70.5
Red Difference	64.5

Table 3: Optimal features for the binary classification using a test frame of 4 samples with $h_r = 2$.

In this study, the optimal features are essentially spectral or temporal. Textural features should give better results on panchromatic images with a spatial resolution of 2m. Indeed, the relatively low spatial resolution (8m) of the Formosat-2 multispectral images turns out insufficient to properly exploit the textural information, particularly of significance in the context of forestry. Besides, the Haralick features were processed on small windows (3×3) which may be insufficient to capture the stand texture. Moreover, in some cases, the broken stems can be oriented in similar directions which might lead to maxima on texture features. Therefore, texture features should be more useful on higher resolution and with an optimization of Haralick parameters.

For the after-storm segmentation process, the best feature is temporal. Indeed, the stand structure (which depends on the tree age) before the storm helps to determine the change degree after the storm. In addition, one can observe that the temporal features lead to higher spectral ranges than the original image bands. Temporal features present a Gaussian distribution, a large spectral range allows to segment the image into different change degrees, whereas the initial image bands present a higher variability, a finer spectral range is necessary to segment subtle changes.

Multiple feature segmentations were also tested. In our experiments, the segmentation performances were better using one feature only. This rather unexpected result can be explained by two reasons. First, the used implementation of the Mean Shift (Comaniciu and Meer, 2002), involves one unique spectral range h_r for multiple features. Indeed, it was initially proposed for gray-scale and color image segmentation. An adaptive spectral range per feature, should enhance the results as it is more appropriate for remote sensing images where spectral distributions of image bands are different. Secondly, the forest canopies are very complex and their variability depends on various parameters or features that are not correlated to the damage degree. For instance, the classical use of the eight initial bands (i.e before and after-storm images) for the segmentation process led to a decrease in the global classification accuracy of 5% (Orny et al., 2010).

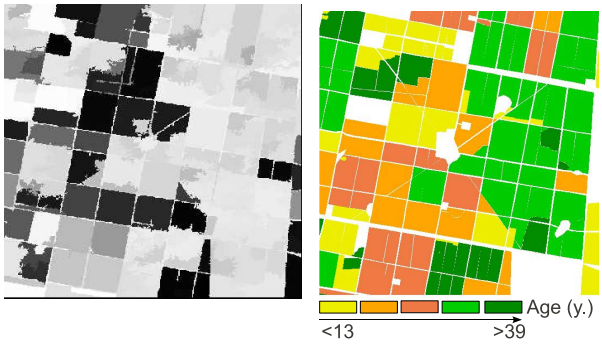
5.3 Segmentation and fragmentation rate

Figure 5 depicts the segmentation results before and after the storm using the respective best features. One can visually distinguish intact and damaged areas. Intact areas are larger and have the same delimitations in both segmentations, whereas the damaged areas are more heterogeneous leading to an over-segmentation into many small regions.

(a) Before-storm segmentation (b) After-storm segmentation

Figure 5: Segmentation of before and after-storm images using the best features, i.e red band and red band ratio respectively.

Figure 6 shows the fragmentation rate (FR) of before-storm segments in gray levels. The lighter the regions, the more damaged

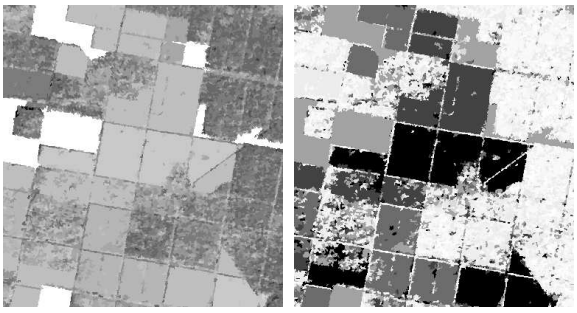


(a) Fragmentation rate (b) Tree stands age map

Figure 6: Comparison between the fragmentation rate and the tree stand ages.

they are. This result matches visually the tree stand age map where the older stands appear to have more damage extent than the younger stands. In fact, among numerous factors, the tree height influences the sensitivity of the tree to the wind. The young stands are dense with small trees which make them more robust to the wind. On the contrary, older stands are less dense, more heterogeneous due to sylvicultural practices, and present higher trees that are more vulnerable and likely to be damaged by storms.

Figure 7 shows the obtained change clusters using the MS spectral classifier. About 30 change clusters are obtained. Given the histogram complexity of this image, it is difficult to separate clusters into two classes automatically. On the contrary, one can observe that after characterizing these regions by their average fragmentation rate, the damaged areas are better discriminated with respect to the intact areas.



(a) MS spectral classification into change classes. (b) Fragmentation rate per change class

Figure 7: MS spectral classification and fragmentation rate

5.4 Map validation

The figure 8 illustrates the final binary map for forest disaster detection with intact and damaged areas. Reference data were collected from orthophotos of 15 cm resolution, that were available after the storm.

The table 4 shows the global confusion matrix, obtained by comparing pixel values between the classification results and the reference data.

	Reference data		
	Intact	Damaged	Total
Intact	1390	487	1877
Damaged	80	2471	2551
Total	1470	2958	4428
Omission error	5.44	16.5	
Commission error	25.9	3.14	
Global accuracy	**87.2**		

Table 4: Confusion matrix for binary classification (Intact/Damaged)

Overall accuracy was of 87.2%. Some small changes or intact areas were not extracted due to the limited spatial resolution of Formosat-2. Besides, in our method, the shadows are not taken into account, therefore the segmentation was not robust to shadow changes.

The INRA [1] inventory data layer of the *Nezer* site references ages of all forest stands. It was confronted to the obtained classification. The overall accuracy and the percentage of damaged pixels (damaged rate) were computed for all available ages. The obtained results are depicted in figure 9.

One can observe that the classification accuracy increases significantly with the class age. The tree height influences its sensitivity to the wind (Cucchi et al., 2005). The damage rate is higher for the older stands and reaches 70% for stands whose age is superior to 25 years.

The detection precision of stands aged from 14 to 39 years (four intermediate classes) is high and ranges between 93.3 and 99.4%.

[1] Institut National de Recherches Agronomiques

Figure 8: Binary map of wind storm damaged areas.

However, the two youngest and the oldest stands have a lower detection rate combined to high omission and commission errors (*cf.* Table 4). The confusion happens between damaged areas and old stands that are heterogeneous and sparse. In addition, small damaged areas or areas with a low intensity damage (leaning trees) are difficult to detect in young dense stands.

These results can be compared to the results of a similar work obtained in (Schwarz et al., 2001) from 10m high resolution multispectral images using a supervised object-oriented approach. Using an automatic object-based approach, our overall accuracy is inferior (87.2%) to the one obtained by (Schwarz et al., 2001) (96%). However, the classification accuracy obtained on trees, aged from 14 to 39 years, is slightly better (96% versus 95%). Good results are obtained on intermediate and older classes. Some problems persist on younger classes as explained above. However, our method requires only a few samples to construct the test frame and only two parameters to tune. Reducing the training and tuning time is essential for emergency mapping.

6 CONCLUSION

We presented, in this paper, an object-based multitemporal change detection method, well-suited for emergency mapping. Our contribution provides two main novelties with respect to similar works

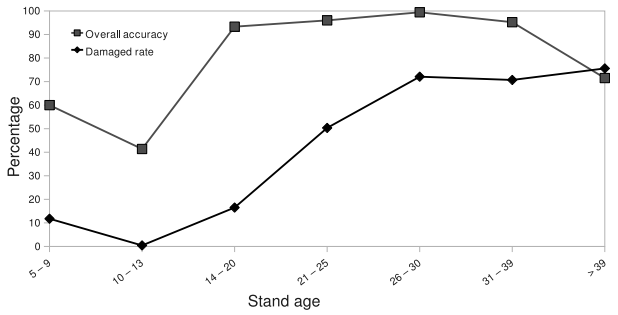


Figure 9: Classification accuracy and damaged pixel rate with respect to age classes.

in forestry. Firstly, an automatic feature selection process, applied to both the segmentation and classification steps, is introduced, whose originality lies on the use of test frames (single or multi-bands) of adequate forest samples. This innovative feature selection process, inspired by camera calibration procedures, allows a rapid evaluation of hundreds of features and combined features. It is applicable to the optimization of any other segmentation algorithm and in the context of any application related to forestry or not. The second originality of our approach is a relevant fragmentation rate, dedicated to storm damages, that ensures an automatic thresholding of the change clusters into a binary map. Our method gives good results. It has the appealing property of requiring only a few samples to construct the test frames, leading to a slight supervision, compared to more traditional supervised methods, hence categorizing it as an *unsupervised* approach. Moreover, our method involves only two user-defined parameters and does not need any statistical assumption, thanks to the powerful Mean-Shift clustering algorithm, at the core of our change detection scheme.

However, the Formosat-2 multispectral images resolution appears to be not well-suited to detect scattered small damages, due to the underlying relatively low spatial resolution (8m). Indeed, this resolution turns out insufficient to properly exploit the textural information, particularly significant in the context of forestry. Panchromatic images, of 2m resolution, have the potential to enhance the current mapping performances. Finally, the confrontation of the resulting binary (intact/damaged) classification to the age class map was consistent and confirmed the increase in sensitivity of the tree to the wind with the age.

Future work will be devoted to the application of our change detection scheme to panchromatic and very high resolution images of the Nezer site. A higher spatial resolution would be certainly more appropriate to capture textural forest details and improve the detection of small changes. A hierarchical approach could also enhance the detection of subtle changes in young dense stands. In addition, in the Mean Shift algorithm, the spectral range h_r should be adapted to each feature, to take advantage of the spectral distribution variability and the combination of multiple features.

7 ACKNOWLEDGMENTS

The authors would like to thank the CNES[2] for their funding through the ORFEO program, the availability of images in the KALIDEOS database (`www.kalideos.cnes.fr`) and the OTB team (`www.orfeo-toolbox.org`) for their technical support.

[2]Centre National d'Etudes Spatiales

REFERENCES

Comaniciu, D. and Meer, P., 2002. Mean shift : A robust approach toward feature space analysis. IEEE Transactions on pattern analysis and machine intelligence 24(5), pp. 603–619.

Conchedda, G., Durieux, L. and Mayaux, P., 2008. An object-based method for mapping and change analysis in mangrove ecosystems. ISPRS Journal of Photogrammetry and Remote Sensing 63(5), pp. 578–589. Theme Issue: Remote Sensing of the Coastal Ecosystems.

Cucchi, V., Meredieu, C., Strokes, A., Coligny, F., Suarez, J. and Gardiner, B., 2005. Modelling the windthrow risk for simulated forest stands of maritime pine (pinus pinaster ait.). Forest Ecology and Management 213, pp. 84–196.

Desclee, B., Bogaert, P. and Defourny, P., 2006. Forest change detection by statistical object-based method. Remote Sensing of Environment 102(1-2), pp. 1–11.

Fraser, R., Abuelgasim, A. and Latifovic, R., 2005. A method for detecting large-scale forest cover change using coarse spatial resolution imagery. Remote Sensing of Environment 95(4), pp. 414–427.

Fukunaga, K. and Hostetler, L., 1975. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory 1, pp. 32–40.

Haralick, R., Shanmugan, K. and Dinstein, I., 1973. Texture features for image classification. IEEE Transactions Systems, Man and Cybernetics 3, pp. 610 – 621.

Im, J. and Jensen, J., 2005. A change detection model based on neighborhood correlation image analysis and decision tree classification. Remote Sensing of Environment 99(3), pp. 326–340.

Inglada, J., 2001. Etat de l'art en détection de changements sur les images de télédétection. Technical Report 2001-23, Cnes.

Kayitakire, F., Hamel, C. and Defourny, P., 2006. Retrieving forest structure variables based on image texture analysis and ikonos-2 imagery. Remote Sensing of Environment 102, pp. 390–401.

Orny, C., Chehata, N., Boukir, S. and Guyon, D., 2010. Characterization of maritime pine forest structure changes with vhr satellite imagery: application to the 24th january 2009 windfall damages cartography. Technical report, CNES.

Otsu, N., 1979. A threshold selection method from gray-level histogram. IEEE Transactions on Systems, Man and Cybernetics 9, pp. 62–66.

Ruiz, L. A., Fdez-sarra, A. and Recio, J. A., 2004. Texture feature extraction for classification of remote sensing data using wavelet decomposition: a comparative study. In: International Archives of Photogrammetry and Remote Sensing. Vol.XXXV, pp. 1682–1750.

Schwarz, M., Steinmeier, C. and Waser, L., 2001. Detection of storm losses in alpine forest areas by different methodic approaches using high resolution satellite data. In: Proceedings of the 21st EARSeL Symposium: Observing our Environment from Space: New Solutions for a New Millenium, pp. 251–257.

St-Louis, V., Pidgeon, A., Radeloff, V., T.J.Hawbaker and Clayton, M., 2006. High-resolution image texture as a predictor of bird species richness. Remote Sensing of Environment 105, pp. 299–312.

Trias-Sanz, R., Stamon, G. and Louchet, J., 2008. Using colour, texture, and hierarchical segmentation for high-resolution remote-sensing. ISPRS Journal of Photogrammetry and Remote Sensing 63, pp. 156–168.

Tuominen, S. and Pekkarinen, A., 2005. Performance of different spectral and textural aerial photograph features in multi-source forest inventory. Remote Sensing of Environment 94, pp. 256–268.

STATISTICAL UNBIASED BACKGROUND MODELING FOR MOVING PLATFORMS

M. Kirchhof[1], U. Stilla[2]

[1] Herborn, Germany –
michael.kirchhof@gmx.de
[2] Photogrammetry and Remote Sensing, Technische Universitaet Muenchen (TUM), Germany –
stilla@tum.de

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: statistical background modelling, detection of moving objects, pixel process

ABSTRACT:

Statistical background modeling is a standard technique for the detection of moving objects in a static scene. Nevertheless, the state-of-the-art approaches have several lacks for short sequences or quasi-stationary scenes. Quasi-static means that the ego-motion of the sensor is compensated by image processing. Our focus of attention goes back to the modeling of the pixel process, as it was introduced by Stauffer and Grimson. For quasi-stationary scenes the assignment of a pixel to an origin is uncertain. This assignment is an independent random process that contributes to the gray value. Since the typical update schemes are biased we introduce a novel update scheme based on the join mean and join variance of two independent distributions. The presented method can be seen as an update for the initial guess for more sophisticated algorithms that optimize the spatial distribution.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

A SCHEME FOR THE DETECTION AND TRACKING OF PEOPLE TUNED FOR AERIAL IMAGE SEQUENCES

F. Schmidt, S. Hinz

Institute of Photogrammetry and Remote Sensing (IPF), Karlsruhe Institute of Technology (KIT), 76128 Karlsruhe, Germany – (florian.schmidt, stefan.hinz)@kit.edu

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: aerial image sequences, object detection, classifier training, people tracking

ABSTRACT:

This paper addresses the problem of detecting and tracking a large number of individuals in aerial image sequences that have been taken from high altitude. We propose a method which can handle the numerous challenges that are associated with this task and demonstrate its quality on several test sequences. Moreover this paper contains several contributions to improve object detection and tracking in other domains, too. We show how to build an effective object detector in a flexible way which incorporates the shadow of an object and enhanced features for shape and color. Furthermore the performance of the detector is boosted by an improved way to collect background samples for the classifier training. At last we describe a tracking-by-detection method that can handle frequent misses and a very large number of similar objects.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

EVENT DETECTION BASED ON A PEDESTRIAN INTERACTION GRAPH USING HIDDEN MARKOV MODELS

F. Burkert, M. Butenuth

Technische Universitaet Muenchen (TUM), Remote Sensing Technology, Arcisstr. 21, 80333 Munich, Germany – (florian.burkert, matthias.butenuth)@bv.tum.de

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: event detection, pedestrians, trajectory interpretation, Hidden Markov Model, aerial image sequences

ABSTRACT:

In this paper, we present a new approach for event detection of pedestrian interaction in crowded and cluttered scenes. Existing work is focused on the detection of an abnormal event in general or on the detection of specific simple events incorporating only up to two trajectories. In our approach, event detection in large groups of pedestrians is performed by exploiting motion interaction between pairs of pedestrians in a graph-based framework. Event detection is done by analyzing the temporal behaviour of the motion interaction with Hidden Markov Models (HMM). In addition, temporarily unsteady edges in the graph can be compensated by a HMM buffer which internally continues the HMM analysis even if the representing pedestrians depart from each other awhile. Experimental results show the capability of our graph-based approach for event detection by means of an image sequence in which pedestrians approach a soccer stadium.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

TRAJECTORY EXTRACTION AND DENSITY ANALYSIS OF INTERSECTING PEDESTRIAN FLOWS FROM VIDEO RECORDINGS

M. Plaue, M. Chen, G. Bärwolff, H. Schwandt

Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany –
(plaue, minjie.chen, baerwolf, schwandt)@math.tu-berlin.de

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: density estimation, human crowd analysis, intersecting pedestrian flows, velocimetry, video analysis

ABSTRACT:

Empirical data of human crowd behaviors are indispensable for the further understanding of pedestrian dynamics. In this paper, we describe a technique for the semi-automatic extraction of pedestrian trajectories from video recordings of human crowds. This method works on data obtained from an arbitrary observation angle and does not require additional information like the heights of the pedestrians etc. It is thus suitable for the analysis of data that have not been specifically prepared for this purpose, such as surveillance videos. We employ this method to analyze video recordings from a series of experiments that we conducted last year to reproduce pedestrian flows under controlled conditions. From these data we also estimate the continuous density of these pedestrian flows via a nearest-neighbor kernel density method which we argue is particularly suited for particle densities in general and human crowds consisting of multiple populations in particular.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

MEASUREMENT ACCURACY OF CENTER LOCATION OF A CIRCLE BY CENTROID METHOD

R. Matsuoka[1,2], N. Shirai[1], K. Asonuma[1], M. Sone[2], N. Sudo[2], H. Yokotsuka[2]

[1] Kokusai Kogyo Co. Ltd., 2-24-1 Harumi-cho, Fuchu, Tokyo 183-0057 Japan –
(ryuji_matsuoka, naoki_shiarai, kazuyoshi_asonuma)@kkc.co.jp
[2] Tokai University Research & Information Center, 2-28-4 Tomigaya, Shibuya-ku, Tokyo 151-0063, Japan –
(ryuji, sone3)@yoyogi.ycc.u-tokai.ac.jp, (sdo, yoko)@keyaki.cc.u-tokai.ac.jp

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: analysis, measurement, accuracy, circle, center, digitization

ABSTRACT:

This paper reports an investigation of the effect of digitization on the measurement accuracy of the center location of a circle by a centroid method. Although general expressions representing the measurement accuracy of the center location of a circle by the centroid method are unable to be obtained analytically, we have succeeded in obtaining the variances V of measurement errors for 39 quantization bits n ranging from one to infinity by numerical integration. We have succeeded in obtaining the effective approximation formulae of V as a function of the diameter d of the circle for any n as well. The results show that V would oscillate on an approximate one-pixel cycle in d for any n and decrease as n increases. The differences of V among the different n would be negligible when $n \geq 6$. Some behaviors of V with an increase in n are demonstrated.

This contribution was selected in a double blind review process to be published within the ***Lecture Notes in Computer Science*** series (Springer-Verlag, Heidelberg).

Photogrammetric Image Analysis

Volume Editors: Stilla U, Rottensteiner F, Mayer H, Jutzi B, Butenuth M
LNCS Volume: 6952
Series Editors: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Weikum G
ISSN: 0302-9743

The article is accessible online through www.springerlink.com.

MULTISCALE HAAR TRANSFORM FOR BLUR ESTIMATION FROM A SET OF IMAGES

Lâmân Lelégard, Bruno Vallet, Mathieu Brédif

Université Paris Est, IGN, Laboratoire MATIS
4, Av. Pasteur
94165 Saint Mandé Cedex, FRANCE
laman.lelegard@ign.fr (corresponding author), bruno.vallet@ign.fr, mathieu.bredif@ign.fr
http://recherche.ign.fr/labos/matis

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: Sharpness, Haar transform, multiscale, calibration

ABSTRACT:

This paper proposes a method to estimate the local sharpness of an optical system through the wavelet-based analysis of a large set of images it acquired. Assuming a space-invariant distribution of image features, such as in the aerial photography context, the proposed approach produces a sharpness map of the imaging device over 16x16 pixels blocks that enables, for instance, the detection of optical defects and the qualification of the mosaicking of multiple sensor images into a larger composite image. The proposed analysis is based on accumulating of the edge maps corresponding to the first levels of the Haar Transform of each image of the dataset, following the intuition that statistically, each pixel will see the same image structures. We propose a calibration method to transform these accumulated edge maps into a sharpness map by approximating the local PSF (Point Spread Function) with a Gaussian blur.

1 INTRODUCTION

Characterizing the spatial resolution of an imaging system is an important field of image processing and is used for assessing its image quality and for restoration purposes.

This characterization can be obtained by shooting some perfectly known objects, preferably periodic patterns such as Foucault resolution targets or Siemens stars (Fleury and Mathieu, 1956) to deduce the smallest periodic detail discernible by the system through the determination of a Modular Transfer Function (MTF) (Becker et al., 2007). This calibration technique is mainly used as a global Point Spread Function (PSF) characterization of the imaging system. However, some imaging systems (mounted with fisheye lenses for instance) show a very space-dependent resolution. In these circumstances, a local study is more suitable and can be done by using a wall of targets, such as Siemens stars (Kedzierski, 2008). Another calibration method consists in materializing a punctual source by a laser beam in order to calculate the PSF (Du and Voss, 2004). However, these approaches require that the optical device go through a calibration procedure in a controlled environment where the appropriate targets are displayed.

Conversely, some blind estimation methods were recently presented that use the edges present in an image. In the case of airborne imagery, a mission over an urban area provides images with a large amount of edges. Assuming a Gaussian PSF and an equal distribution of edge orientations, (Luxen and Forstner, 2002) estimates the standard deviation of the Gaussian blur. An alternative way of considering the problem is to study the image frequencies by comparing the local spectrum (obtained after integrating the Fourier Transform of a local patch in an image across the polar coordinate theta) to the global image spectrum (Liu et al., 2008). By varying the size of the patch used, one can choose a compromise between the quality of the frequency estimation and of the localization.

In an intermediate approach (Zhang and Bergholm, 1997) local information (like edges) is considered at different scales using differences of Gaussians. One interesting assessment is the behavior of edges according to the scale at which they are observed. Sharp edges vanish at coarse scale whereas diffuse ones become sharper when looked at coarser scales. An application proposed by (Zhang and Bergholm, 1997) is blur estimation applied to deduce depth from focus. Multi-scale analysis is also an interesting compromise between spatial and frequency accuracy. The use of Haar wavelets in (Tong et al., 2004) is an interesting intermediate solution. This is the framework that we investigate in this work.

The blind estimation methods cited above rely on a single image so the blur caused by the optical system cannot be distinguished from the smoothness of the imaged object itself. Our contribution is twofold: we overcomes this limitation by extending Tong's method (designed for a single image) to a large set of images, and we propose a quantitative characterization of sharpness through a blur radius.

2 OVERVIEW

Our method is based on Tong's blur detection method that relies itself on Haar wavelets. We will start by recapitulating his approach, then explain the two contributions of our paper, and finally expose the assumptions that we make on our datasets.

2.1 Haar wavelets

Tong's method comes down to three main steps (Figure 1):

1. Do a 3 levels Haar wavelets transform.
2. Extract multi-scale normal edges maps E_l^{norm} and maximal edge maps E_l^{max}
3. Apply rules to these maps to estimate the sharpness.

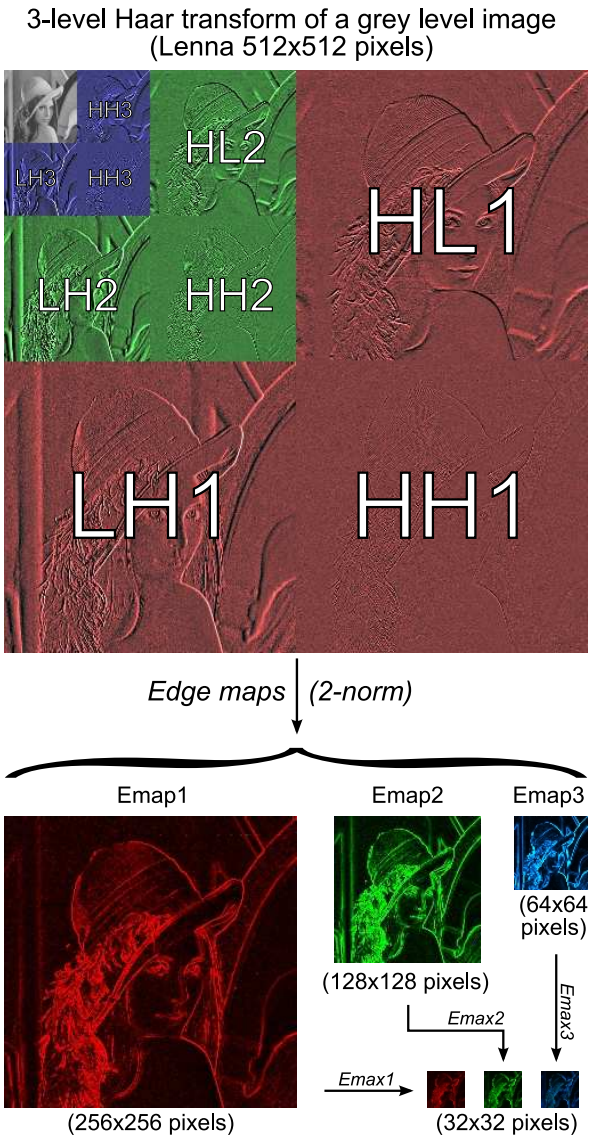


Figure 1: Haar wavelets transform and edge maps

The Haar decomposition of an image I is defined by:

$$\begin{pmatrix} LL_{l+1}(i,j) \\ LH_{l+1}(i,j) \\ HL_{l+1}(i,j) \\ HH_{l+1}(i,j) \end{pmatrix} = M_{Haar} \begin{pmatrix} LL_l(2i,2j) \\ LL_l(2i,2j+1) \\ LL_l(2i+1,2j) \\ LL_l(2i+1,2j+1) \end{pmatrix} \quad (1)$$

where L and H stand for Low and High frequencies, $LL_0 = I$ and the Haar matrix is:

$$M_{Haar} = \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \quad (2)$$

For each level $l = 1..3$, an edge map is obtained by calculating (for each pixel) the norm:

$$E_l^{norm} = \sqrt{LH_l^2 + HL_l^2 + HH_l^2} \quad (3)$$

The E_l^{norm} do not have the same size, so Tong proposes to define a *maximal* edge map E_l^{max} of constant size to make possible a comparison between different levels:

$$E_l^{max}(i,j) = \max_{di,dj=0}^{2^{4-l}-1} E_l^{norm}(2^{4-l}i + di, 2^{4-l}j + dj) \quad (4)$$

thus all E_l^{max} are $2^4 = 16$ times smaller than the input image. The E_l^{max} measure the level of detail of the image at scale 2^l on $16x16$ pixels blocks. Tong et al. choose to apply rules based on inequalities on the E_l^{max} in order to characterize qualitatively the image sharpness.

2.2 Our approach

The novelty introduced in this paper compared to Tong's approach is twofold:

1. Compute an average $\bar{E}_l^{max}$ of the E_l^{max} over a large set of images acquired with the same imaging system, such that $\bar{E}_l^{max}$ characterize only the optical quality of the imaging system itself. Obviously, $\bar{E}_l^{max}$ will also depend on the statistical properties of the set of images used.

2. Exploit the $\bar{E}_l^{max}$ to define a quantitative measure of the local sharpness. We chose to quantify local sharpness by assimilating the PSF to a Gaussian blur which radius (σ = standard deviation) we will estimate. In other terms, we look for σ as a function:

$$\sigma = \sigma(\bar{E}_1^{max}, \bar{E}_2^{max}, \bar{E}_3^{max}) \quad \sigma : \mathbb{R}^{+3} \to \mathbb{R} \quad (5)$$

The main idea developed in this paper is to look for $\sigma(...)$ as the composition of two functions:

$$\sigma = c(r(\bar{E}_1^{max}, \bar{E}_2^{max}, \bar{E}_3^{max})) \quad (6)$$

- $r : \mathbb{R}^{+3} \to \mathbb{R}$ is a space reduction function, which will reduce our problem from 3 to 1 dimensions. We will explain what properties this function should have and propose a pertinent choice for this function in the next section.

- $c : [0,1] \to \mathbb{R}$ is a monotonous *calibration* function linking an r value to an actual blur radius σ. Because the $\bar{E}^{max}$ depend on the actual statistical properties of the dataset used, a calibration function c should be defined for each dataset. Computation of this calibration function as well as its dependence on image statistics is studied in the next section.

2.3 Assumptions

Characterizing an optical system based on a set of images that it acquired will only be valid statistically if these images have good statistical properties. In particular, the following assumptions should be made on the image dataset:

- Camera settings are constant for all the images.
- Shot objects must be in focus.
- Images should be shot without motion blur.
- The edge presence probability is uniform on the whole image.
- The number of images should be large enough.

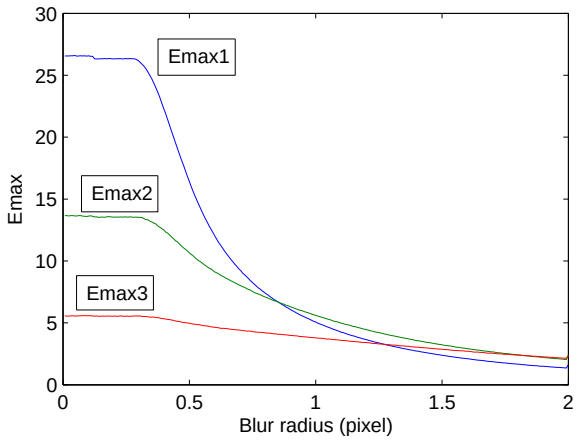


Figure 2: $\bar{E}^{max}$ values for white noise images blurred by a Gaussian of increasing radius.

All these assumptions are usually verified in the case of aerial imagery, provided that motion blur is corrected (by charge transfer for instance), and that the area covered has sufficient texture and edges (forest, city, ...) It is not the case for landscape photographs where the edges are localized at the bottom of the images, and objects at various distances cannot all be in focus simultaneously.

3 BLUR ESTIMATION FROM EDGE MAPS

The aim of this section is to choose the space reduction function r, and to propose a method to compute the calibration function c from a dataset.

3.1 A first experiment

In order to make the exposition clearer, we will start by a simple experiment to exhibit the dependence of the $\bar{E}^{max}$ on blur. The experiment consists in computing the $\bar{E}^{max}$ on a dataset of white noise images (with a given standard deviation), blurred by a Gaussian blur of known radius σ varying from 0 to 2 pixels (Fig. 2). It is easy to see that $\bar{E}^{max}$ depends linearly on the contrast, so their absolute value do not convey a real meaning. Interestingly,the $\bar{E}_1^{max}$ and $\bar{E}_2^{max}$ curves cross between 0.5 and 1, and $\bar{E}_2^{max}$ and $\bar{E}_3^{max}$ cross between 1 and 2. This is quite natural as $\bar{E}_1^{max}$, $\bar{E}_2^{max}$ and $\bar{E}_3^{max}$ corresponds to scales 1, 2 and 4 pixels respectively, and a Gaussian of standard deviation σ is a structure of size (diameter) 2σ. This is very coherent with the idea that $\bar{E}_i^{max}$ exhibits certain scales in the image. Quite logically, the curves are flat below $\sigma = 0.5$ pixels as structures of size lower than 1 pixel are not representable, such that blurs of radii lower than 0.5 are not discernible.

3.2 Space reduction function

The main interest of the space reduction function r is to make calibration easier without loss of generality. In particular, we can make r invariant to contrast by expressing it:

$$r(\bar{E}_1^{max}, \bar{E}_2^{max}, \bar{E}_3^{max}) = s(E_1, E_2)$$

$$\text{with} \quad E_1 = \frac{\bar{E}_2^{max}}{\bar{E}_1^{max}} \quad E_2 = \frac{\bar{E}_3^{max}}{\bar{E}_2^{max}} \tag{7}$$

The $\bar{E}^{max}$ ratios are displayed in Fig. 3 and show quite similar behavior but at a different scale. Thus we can simply choose

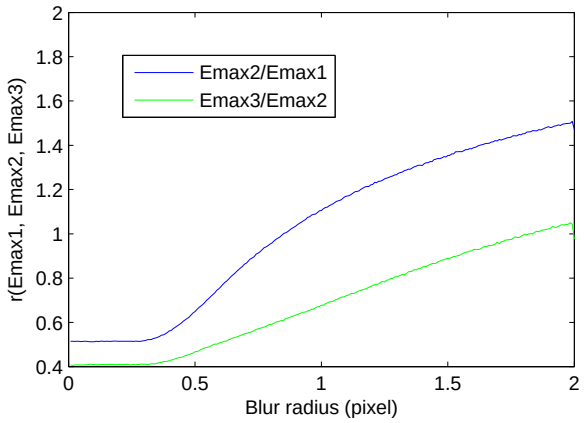


Figure 3: Ratios of $\bar{E}^{max}$ values for white noise images.

the ratio corresponding to our scale of interest, or if we are interested in multiple scales, average the corresponding ratios. If scales larger than 4 pixels are of interest, then we can compute more $\bar{E}_{i+1}^{max}/\bar{E}_i^{max}$. In our case, we are interested in characterizing optical systems, that usually have blurs of radii smaller than one pixel, such that we will simply choose:

$$r(\bar{E}_1^{max}, \bar{E}_2^{max}, \bar{E}_3^{max}) = E_1 = \frac{\bar{E}_2^{max}}{\bar{E}_1^{max}} \tag{8}$$

3.3 Calibration function

As the function r is monotonous in σ in our experiment (except for blur radii below 0.4 pixels that are indiscernible because they reach the Shannon limit of the sensor), this relation ship can be inversed to get σ as a function of r. This is exactly the definition of our calibration function c. Consequently, if we had a ground truth (a perfectly sharp image corresponding to each image acquired by the optical system that we want to characterize), then we could compute $r(\ldots)$ for these perfectly sharp images blurred with various blur radii σ, and get our calibration function c as the inverse of this function.

In order to compute calibration functions in real cases, we will exploit the idea that the statistics of natural images are relatively insensitive to scaling. More precisely, given a dataset of images acquired with a given optical system, we will build a dataset of perfectly sharp images by subsampling with a factor greater than the largest expected blur (a factor 4 gives this guarantee in most cases). We will assume that this dataset has statistical properties close to those of the perfect dataset at full scale (that we cannot have), and compute our calibration function on that subsampled dataset.

The aim of the next subsection is to estimate the validity of the assumption that we make in our approach that our datasets have scale invariant statistical properties.

3.4 Sensitivity to scale

The statistical properties of natural images are complex and have been widely studied. Some work seem to show that invariance to scale is only true at lower scales (Huang and Mumford, 1999). Moreover, the dataset that we use to estimate the blur of our optical system is composed of aerial images that have specific properties that might differ from those of natural images in general. Thus we chose to test our assumption that r is relatively invariant to scale, by a simple experiment: we evaluated the calibration functions for an input dataset with various subsampling factors.

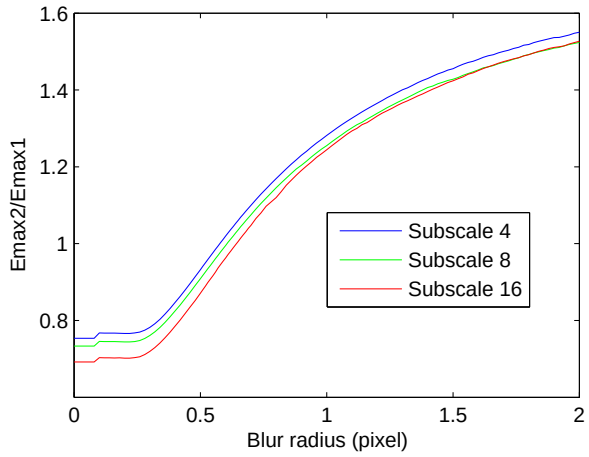


Figure 4: Ratios of $\bar{E}^{max}$ values for subsampled aerial images.

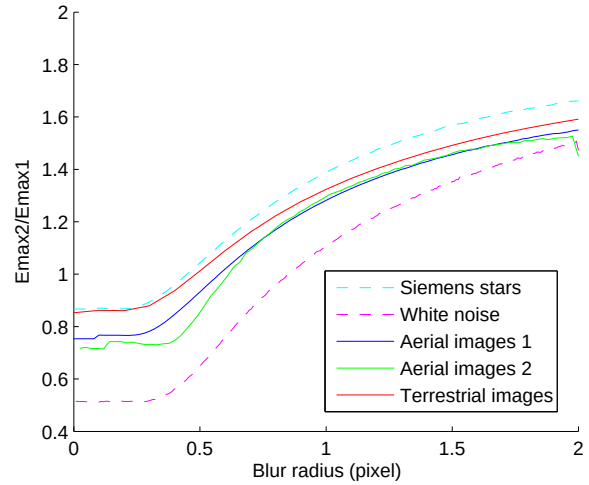


Figure 5: Ratios of $\bar{E}^{max}$ values for synthetic (dashed lines) and real (solid lines) images.

The result of this experiment is displayed in Fig. 4. We see on those curves that the error made by using the calibration function at scale 4 instead of 16 will lead to an error around 0.1 pixel on the blur radius estimation in our zone of interest (0.4 to 1 pixel). This means that we can expect this precision when using the calibration curve at scale 4 to approximate the (unknown) calibration curve at scale 1. This is clearly a limitation to our approach as 0.1 pixel is a rather high error for blurs ranging from 0.4 to 1 pixel. However, we can notice that except in the area below 0.4 pixels where blur cannot be distinguished, the blur radii between different scales are proportional. This means that even if the absolute precision is poor, the relative precision is good, such that our approach is pertinent to locate flaws in the optical system as areas where the blur radius increases. In other terms, the limitation on the precision of the estimation will not impair the interpretation of the result.

A last point of interest is to understand the influence of the image dataset statistics on the calibration curve.

3.5 Sensitivity to image statistics

We have already built the calibration function for noise images and for subsampled aerial images, which have rather different statistical properties. In particular, aerial images present structures

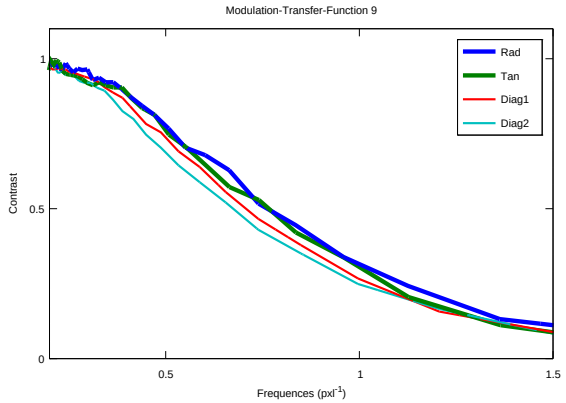


Figure 6: Point spread function at one of the 9 points estimated using a Siemens star.

at various different sizes whereas noise images have mostly structures of sizes close to a pixel. To complete the comparison, we also built calibration functions for a second aerial image dataset, another dataset coming from terrestrial mobile mapping, and a set of synthetic images of packed Siemens stars. The results are displayed in Fig. 5.

We first note that our two synthetic datasets (white noise and Siemens stars) have the most extreme values, one displaying the most irregular structures (white noise), and the other the most regular (Siemens stars). The real datasets are between those extremes: terrestrial imagery in urban areas usually displays large structures, so it is closer to the structured Siemens stars calibration function. Aerial image dataset are quite intermediate, and show similar behavior except around 0.5 pixels. This probably comes from the fact that the second dataset contains more forests which brings more details at very low scales.

In conclusion, the calibration curves are close enough on real images to make visualization of the flaws of the imaging system quite independent of the curve used. For this application, an "average" calibration curve can be used, which saves a lot of computation time. Computing a calibration curve requires to subsample each image of the dataset then compute the E^{max} over each subsampled image, which roughly takes one hour for one thousand images.

3.6 Comparison with blur estimation using Siemens stars

A classical approach to estimate the quality of an optical system is to compute its Modulation Transfer Function (MTF). This can be done by acquiring an image of a Siemens star, then estimating the contrast at various distances of the center (corresponding to a spatial frequency) and in various directions (usually 4). The first aerial dataset that we used contains such a Siemens star that is visible in 9 images, so we applied this procedure to estimate the MTF at 9 different points of the imaging system. The result for one of the points is displayed in Figure 6. We notice that the curves in the various directions are very close, showing isotropy. We estimated a blur radius from these curves for the nine points by computing the contrast as a function of the blur radius and inversing the relationship. The results are displayed in Fig 7. As expected, the error is below 0.1 pixel, and the blur radius is slightly exaggerated by our approach.

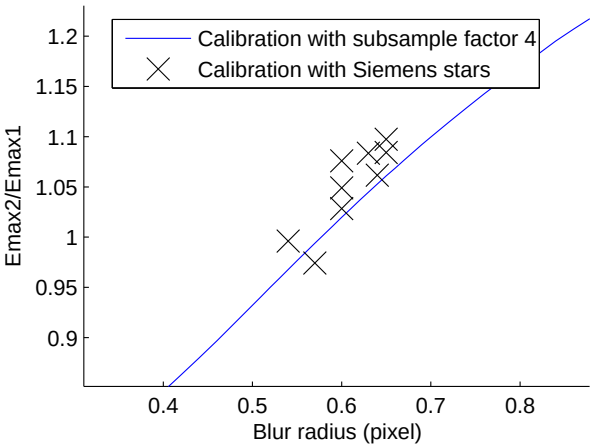


Figure 7: Comparison with blur estimation using Siemens stars: black crosses are at coordinate $(\sigma_{Siemens}, r)$ where $\sigma_{Siemens}$ is the blur estimation with Siemens stars, and $r = \bar{E}_2^{max}/\bar{E}_1^{max}$ is read at the Siemens star center using our approach. The difference between blur estimation using the two approaches is given by horizontal distance from crosses to the calibration curve (blue).

4 RESULTS AND DISCUSSION

Based on the proposed methodology, we are now able to build blur radius maps, with the limitations quoted above, for any optical system for which we have an image dataset satisfying the assumptions of Section 2.3. We will now display the results of this methodology applied to aerial and terrestrial imagery.

4.1 Experiment on aerial imagery

A first set of image is provided from an airborne photogrammetric mission with a DMC. The results of our sharpness estimation for these images are displayed in Figure 8. The small number of images (157) for this mission is compensated by the large resolution of the images (7680x13824) which is a mosaic of four individual panchromatic images. On these images, the top and the bottom areas of the imaging system have a lesser resolution than the center: this could be interpreted as an effect of the deformation (projection) applied to the four images during mosaicking. The second visible artifact is the vertical line in the middle of the figure: the decrease in sharpness in this part of the imaging system may come from the seam between the left and right images. We also notice that some small area (such as the one at the top left) seem more blurry than the average: this might be interpreted as flaw in (or a dust on) the lens or the sensor. Yet it should be noticed that the blur radius is always less than a pixel.

4.2 Experiment on streetside imagery

A second experiment was led on streetside urban images obtained by a camera mounted on a mobile vehicle. The distortion of the camera was corrected prior to applying our method, so the sharpness of the entire pipeline is estimated (Figure 9). One can notice that the grid used for distortion correction is perfectly visible. The sharper area at the bottom of the image is probably due to the fact that our assumption on the homogeneity of edge distribution is not verified as this part of the image always sees the road.

In conclusion, our approach allows us not only to evaluate the quality of an optical system, but also to detect if the image underwent alterations such as interpolations.

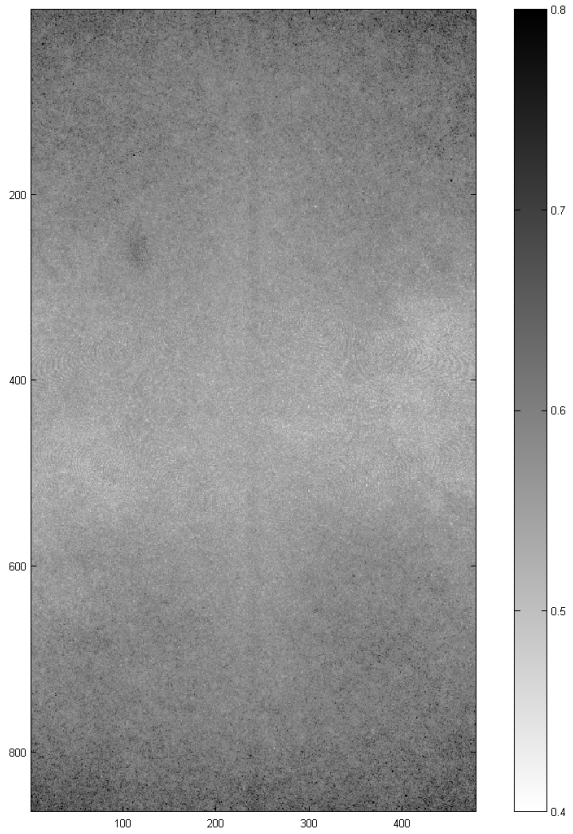


Figure 8: Sharpness image obtained through calibration for the airborne experiment

5 CONCLUSION AND PERSPECTIVES

Our work aims at quantifying the amount of blur induced by an optical system from a large set of images that it has acquired. This is extremely useful in an operational context as it avoids immobilizing an expensive resource (an aerial camera) in a lab to perform its evaluation. It can also be used as a complement to lab calibration as manipulations of the imaging system to transfer it from the lab to the plane might slightly alter its characteristics. Finally, we can think of another application of our method that would be to evaluate the stability of the optical quality in time during its utilization, enabling online certification of the imaging system.

We have shown that our approach allows to build a sharpness map of the imaging system, such that our sharpness estimation is much denser than sparse approaches based on Siemens stars for instance. We have shown that our estimation has an absolute accuracy around 0.1 pixel (in blur radius) which is close to what can be achieved based on MTF estimation. But more importantly, we have a very good relative quality which allows for an easy visual inspection of the localization of possible quality artifacts in the image.

The method developed is targeted at a very limited blur radius range of interest, but can be easily extended by using other ratios for dimension reduction, and using more than 3 Haar levels.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Didier Boldo for the original idea that gave birth to this work.

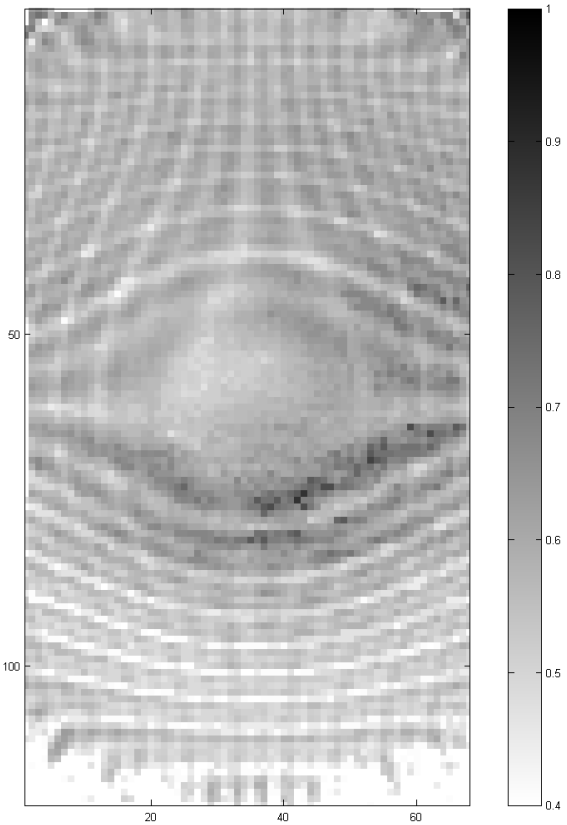


Figure 9: The sharpness estimate for streetside experiment exhibits the grid used for the correction of the distortion

REFERENCES

Becker, S., Haala, N., Honkavaara, E. and Markelin, L., 2007. Image restoration for resolution improvement of digital aerial images : A comparison of large format digital cameras. In: SFPT (ed.), ISPRS Commission Technique I. Symposium, Marne-la-Vallée, France, pp. 5–10.

Du, H. and Voss, K., 2004. Effects of point-spread function on calibration and radiometric accuracy of ccd camera. In: Applied Optics, Vol. 43number 3, pp. 665–670.

Fleury, P. and Mathieu, J., 1956. Chapitre: 18 - photographie projection. In: Eyrolle (ed.), Image Optique, Paris, France, pp. 413–432.

Huang, J. and Mumford, D., 1999. Statistics of natural images and models. In: IEE Conf. on Computer Vision and Pattern Recognition, pp. 541–547.

Kedzierski, M., 2008. Precise determination of fisheye lens resolution. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 37, Beijing, China, pp. 761–764.

Liu, R., Li, Z. and Jia, J., 2008. Image partial blur detection and classification. In: IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, pp. 1–8.

Luxen, M. and Forstner, W., 2002. Characterizing image quality: Blind estimation of the point spread function from a single image. In: Photogrammetric Computer Vision, Graz, Austria, pp. 211–217.

Tong, H., Li, M., Zhang, H. and Zhang, C., 2004. Blur detection for digital images using wavelet transform. In: IEEE International Conference on Multimedia and Expo, Vol. 1, Taipei, Taiwan, pp. 17–20.

Zhang, W. and Bergholm, F., 1997. Multi-scale blur estimation and edge type classification for scene analysis. In: International Journal of Computer Vision, Vol. 24, pp. 219–250.

REFLECTANCE ESTIMATION FROM URBAN TERRESTRIAL IMAGES: VALIDATION OF A SYMBOLIC RAY-TRACING METHOD ON SYNTHETIC DATA

Fabien Coubard [a], Mathieu Brédif [a], Nicolas Paparoditis [a], Xavier Briottet [b]

[a] MATIS, Institut Géographique National, 73 avenue de Paris F-94160 SAINT MANDÉ, France
(fabien.coubard, mathieu.bredif, nicolas.paparoditis)@ign.fr
[b] DOTA, ONERA, 2 avenue Edouard Belin 31055 TOULOUSE, France

Working Groups III/4, III/5

KEY WORDS: radiometry, reflectance, BRDF, ray-tracing, terrestrial images

ABSTRACT:

Terrestrial geolocalized images are nowadays widely used on the Internet, mainly in urban areas, through immersion services such as Google Street View. On the long run, we seek to enhance the visualization of these images; for that purpose, radiometric corrections must be performed to free them from illumination conditions at the time of acquisition. Given the simultaneously acquired 3D geometric model of the scene with LIDAR or vision techniques, we face an inverse problem where the illumination and the geometry of the scene are known and the reflectance of the scene is to be estimated. Our main contribution is the introduction of a symbolic ray-tracing rendering to generate parametric images, for quick evaluation and comparison with the acquired images. The proposed approach is then based on an iterative estimation of the reflectance parameters of the materials, using a single rendering pre-processing. We validate the method on synthetic data with linear BRDF models and discuss the limitations of the proposed approach with more general non-linear BRDF models.

1 INTRODUCTION

1.1 Context

The development of digital cameras has led to new possibilities for radiometric processes: the linear response of the sensor with respect to incoming radiance gives radiometer abilities to the camera. So understanding the physical processes that lead to the formation of digital images is a key to developing many applications, for professionals as well as for the general public. For instance, terrestrial images are nowadays widely available on the Internet: Google Street View in 2007, Microsoft Bing Streetside in 2009 (North America only), or the research project iTowns (Devaux and Paparoditis, 2010). These images have very high spatial resolution (a few centimeters); so the radiative phenomena must be modelized with the same resolution. In this work, we focus on images acquired from terrestrial vehicles, with large spectral bands in the visible domain. The seamless visualization of such images has to deal with the dependence of the pixel values on the illumination conditions at the time of acquisition. For instance, the illuminated and shadowed areas are determined by that time, and can be a disturbance for the viewer. On the long run, the objective is to free images from their illumination conditions, enabling applications such as relighting or augmented reality. That can be done through the estimation of the intrinsic color properties of the scene surface: the reflectance.

1.2 Related work

The estimation of one or several radiometric parameters of a scene, from a set of observations (mainly images), is referred to as inverse rendering in computer graphics. Patow and Pueyo (2003) propose a survey of these methods. As part of it, inverse reflectometry aims at retrieving reflectance properties of objects from one or several images. The hypotheses and input data of these works can be very different: global or only direct illumination, controlled or uncontrolled light sources, invariant or textured reflectance map, single or multi-angular. Machida et al. (2007) propose an inverse method using photon mapping, with a parametric model of BRDF (Bidirectional Reflectance Distribution Function), from images with very directional light source with known position. Yu et al. (1999) use inverse radiosity in a multi-scale hierarchical method, from a large set of images from different camera positions. Boivin and Gagalowicz (2001) estimate BRDF from a single image. Lensch et al. (2003) estimate spatially-varying BRDF without considering reflections between objects. The main goal of these methods is often to produce augmented reality from real images, such as addition of new objects or relighting under user-defined lighting conditions.
In the remote sensing community, some authors follow a more physical approach, looking for accurate estimation of properties, mainly analyzing outdoor aerial or satellite images for classification purposes. A physical modelization of the different radiance terms in urban images has been proposed by Miesch et al. (2000). Using multi-view aerial images, Martinoty (2005) uses the same modelization for classification of roof materials with BRDF parameters as criteria, and Lachérade et al. (2008) estimates the albedo (in the Lambertian case) for accurate synthesis of satellite images under any atmospheric conditions.

1.3 Proposed approach

In this work, the input data is a set of terrestrial images shot from a mobile-mapping vehicle. We also suppose that we have a 3D geometric model of the scene, segmented in areas that are homogeneous in reflectance. That segmentation could be obtained with the LIDAR backscattering data (independent of illumination) simultaneously acquired by the vehicle, but that topic is not tackled in this work (see (Müller et al., 2007) for an example of facade modeling). We also consider the data are acquired under good atmospheric conditions which are supposed to be known through observable variables for aerosol (type, concentration) and gas (water vapour content). The method aims at enabling radiometric corrections of terrestrial images (shadow removal, relighting, etc), and intends to be as general as possible, though adapted to outdoor urban scenes. Furthermore, we do not look for purely image-based methods, such as tonemapping; but

we instead try to use a physical model of the image radiometry, in order to explore the potential of physically-based methods. To summarize, our method uses the following input data: known but uncontrolled light sources (sun, sky), a 3D geometric model of the scene segmented into regions of homogeneous reflectance properties, and single or multi-view calibrated images. It takes into account the reflections between objects and the bidirectional behaviour of materials (up to the limitations discussed in section 4.3). This paper presents the physical background of that study, in section 2, and the simulation method used in section 3. Then the inversion algorithm is presented in section 4, as well as the motivations for symbolic ray-tracing. Results on a synthetic scene are then shown in section 5, as well as the computational limitations of the inversion method.

2 RADIATIVE TRANSFER IN URBAN SCENES

An image is the result of the propagation of light rays from the light sources to the sensor. This propagation of a light ray is affected by the medium it goes through and the surfaces it hits. In the case of an outdoor urban scene: the source is the sun, the medium is the atmosphere, and the surfaces are the objects of the scene: buildings, soil... So the modelization of the radiometric pixel values requires the modelization of the radiative transfer in the atmosphere and the reflection on the objects of the scene.

2.1 Participating medium

The solar light interacts with the atmospheric particles (gas molemolecules or aerosol) located at point Q. Part of its energy is absorbed and scattered, in a proportion given by the attenuation coefficient σ_e. Along a unit vector $\vec{\omega}$, the radiance going out of the particle at wavelength λ is given by the radiative transfer equation (Meyzonnette and Lépine, 1999):

$$L(Q+\delta\vec{\omega},\vec{\omega},\lambda) = L(Q,\vec{v},\lambda) - \sigma_e \left(L(Q,\vec{\omega},\lambda) - J(Q,\vec{\omega},\lambda)\right)\delta \tag{1}$$

where δ is a small length, and J is the source function that takes into account the scattering of the light hitting the particle from other directions than $\vec{\omega}$ and volumetric light emission.

2.2 Interaction with surfaces

When hitting a surface at point P, the light is reflected according to its Bidirectional Reflectance Distribution Function (BRDF), that links the outgoing radiance L_{out} in a given direction of reflection $\vec{\omega}_{out}$ to the radiance distribution L_{in} coming from the incident directions $\vec{\omega}_{in}$ (Nicodemus et al., 1977). Considering a non-emitting surface, and denoting by Ω_P the hemisphere directed by the surface normal $\vec{n}_p$, this relation is (Kajiya, 1986):

$$L_{out}(P,\vec{\omega}_{out},\lambda) = \int_{\vec{\omega}_{in}\in\Omega_P} brdf_P(\vec{\omega}_{in},\vec{\omega}_{out},\lambda)\cdot L_{in}(P,-\vec{\omega}_{in},\lambda)\langle\vec{\omega}_{in}\cdot\vec{n}_p\rangle d\vec{\omega}_{in} \tag{2}$$

The high dimension of the parameter space of BRDF distributions prevents its accurate modelization with a limited number of parameters. Many models exist, each fitted for certain materials; for instance the models of Torrance and Sparrow (1967), Cook and Torrance (1981) and Ward (1992) are widely used in computer graphics (Yu et al., 1999; Machida et al., 2007). For reasons of space-complexity of the inversion method (see section 4.3), and at the expense of physical reliability, we propose to limit our search space to the particular class of kernel BRDFs.

We denote a **kernel BRDF** as a linear combination of parameterless BRDF basis functions $f_i^{seg(P)}(P,\vec{\omega}_{in},\vec{\omega}_{out})$ weighted by geometry-agnostic functions $\rho^{seg(P)}(\lambda)$ which are the unknown constants for each scene surface segment of homogeneous reflectance ($seg(P)$ denotes the index of the scene segment containing P).

$$brdf_P(\vec{\omega}_{in},\vec{\omega}_{out},\lambda) = \sum_i \rho_i^{seg(P)}(\lambda)\ .\ f_i^{seg(P)}(P,\vec{\omega}_{in},\vec{\omega}_{out}) \tag{3}$$

For instance, the following kernel BRDF blends a Lambertian model (isotropy of the outgoing radiance; corresponding to a perfectly diffuse behaviour), weighted by ρ_d, and a perfectly specular model (mirror behaviour), weighted by ρ_s :

$$brdf_P(\vec{\omega}_{in},\vec{\omega}_{out},\lambda) = \frac{\rho_d(\lambda)}{\pi} + \rho_s(\lambda).\delta_{\{bisect(\vec{\omega}_{in},\vec{\omega}_{out})=\vec{n}_P\}} \tag{4}$$

2.3 The specific case of urban terrestrial images

High spatial resolution urban terrestrial imagery imply specific conditions that prevent us from using the hypotheses commonly assumed in aerial and satellite remote sensing. First, the objects of the scene cannot be considered as equally distant from the camera, so the scene cannot be modelized as a plane (as in the radiometric equalization proposed by Chandelier and Martinoty (2009)). This also means that the reflections between objects have to be taken into account. However, the reflected light may be neglected after a few reflections, because its radiance decreases with the number of reflections (due to the geometric attenuation, the transmittance, the reflectance). The importance of reflections decreases also with the square of the distance of the reflecting surface. Lachérade et al. (2008) show that it may be neglected after two reflections.

The Lambertian hypothesis can be valid with low or medium resolution images, because each pixel integrates the contribution of a relatively large surface: the roughness, that is relative to the resolution, usually is high enough to observe an almost perfectly diffuse reflectance. This implies that the BRDF does not depend on the angles of incidence and reflection, so it makes it possible to compute the irradiance on each surface without saving the information of incident angle for each ray. This might be very convenient for further inversion (Lachérade et al., 2008). Unfortunately, the Lambertian hypothesis is no longer valid when the image has a resolution of a few centimeters (Martinoty, 2005).

The transmittance along each rectilinear portion of a ray can be computed simply in the case of a homogeneous medium without scattering, by multiplying the radiance by $e^{-\sigma_e d}$ where σ_e is the attenuation coefficient and d the length of the rectilinear portion. But in general, there is scattering and inhomogeneous medium, and we have to stochastically divert rays while they propagate. That complete computation, though possible, might be very costly. However, a simplification can be adopted: the atmosphere within the convex hull of the 3D model is almost perfectly transparent, due to the low distances between sensor and objects. This hypothesis may be numerically validated; the extinction coefficient σ_e is linked to the horizontal visibility V (that is the greatest distance at which an object can be seen): $e^{-\sigma_e V} = 0.02$. For an atmosphere with a visibility of 23 km, the transmittance at a distance of 30 m is 0.995. So the absorption and the scattering inside the scene may be neglected. Particularly, it implies that fog or fumes at the street level are not taken into account. Furthermore, the size of the scene can be considered as negligible compared to distance at which occurs the atmospheric

scattering; the downward radiance of the sky coming from a given direction can be considered as constant on the whole scene.

3 DIRECT PROBLEM: IMAGE SYNTHESIS

3.1 Ray-tracing

Computing the total radiance analytically is untractable in the general case; we have to use numerical methods to estimate the radiance at the sensor level (Miesch et al., 2000). These rendering methods have been developed for realistic image synthesis; see (Dutre et al., 2006) for a description of the different rendering methods. In this work, we chose ray-tracing method because of its genericity: it can simulate complex BRDF models, and takes into account any scattering or reflection or absorption phenomenon (within the limits of geometrical optics). Furthermore, it allows any kind of geometry, since it only uses the local normals to the surfaces and ray/surface intersection tests; no discretization of surfaces is needed.

The principle of ray-tracing is to use the local equations of radiative transfer (1) and (2) to compute the exitant radiance for every point hit by a light ray. To compute the radiance at the sensor level for a pixel of the image, we throw many rays from the camera in the FOV of the pixel. When hitting a surface, a ray is reflected in a stochastically sampled direction, until it hits a light source. This produces a light path C, made of 3D line segments, which carries a radiance $L(C)$ and a sampling probability $pdf(C)$. The Monte-Carlo estimate (5) gathers the radiances $L(C_k)$ carried by all the paths C_k generated by rays thrown from a pixel FOV into the radiance value of the pixel $L_{sensor}^{synthetic}$, and may even be slightly modified to take into account the point spread function (PSF) of the sensor (Pharr and Humphreys, 2004):

$$\frac{1}{N}\sum_{k=1}^{N}\frac{L(C_k)}{pdf(C_k)} \underset{N\to\infty}{\to} L_{sensor}^{synthetic} \qquad (5)$$

3.2 Implementation details

The criteria of choice of a ray-tracer are: an unbiased estimation, a reasonable computation time (real-time is however out of scope), open source software to allow modifications, a good control through physically pertinent parameters. For these reasons, we chose to use the GPL physically-based renderer LuxRender, that is based on PBRT (Physically-Based Ray-Tracing) (Pharr and Humphreys, 2004).

In order to fasten the convergence of the estimation (5), LuxRender uses importance sampling to drive the stochastic choice of the reflection direction; the BRDF values are used as a probability density function (pdf). When evaluating the radiance coming directly from the light sources, the pdf takes into account the BRDF and the angular distribution of light sources (multiple importance sampling (Veach, 1997)).

To provide an unbiased estimate, LuxRender emulates the sampling of paths with an unbounded number of reflections (using a Russian roulette technique). We however chose to limit the maximum number of allowed reflections N_{max} to 3 to prevent the combinatorial explosion of our technique, driven by the observation in section 2.3 that the reflected light may be neglected after a few interactions.

As discussed in section 2.3, the downward radiance from the sky can be represented as a function of the direction. An environment map is thus sufficient to describe the radiance coming from the sky, i.e. a Look-Up Table (LUT) giving for each direction the value of the downward radiance from the sky (figure 1). That downward radiance can be computed by a radiative transfer code, such as 6S (Vermote et al., 2002). That environment map and the irradiance at top of atmosphere and the optical thickness of the atmosphere are sufficient to model the light sources and the participating medium in our case.

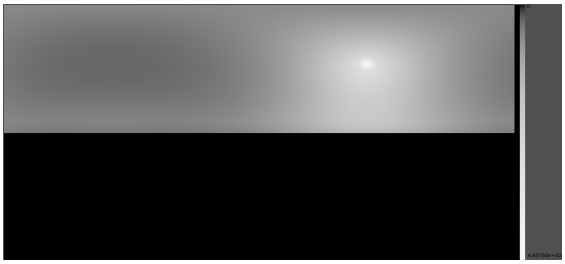


Figure 1: An example of environment map computed with 6S, at wavelength $550nm$. The solar halo is visible, but this environment map does not represent the direct solar radiance.

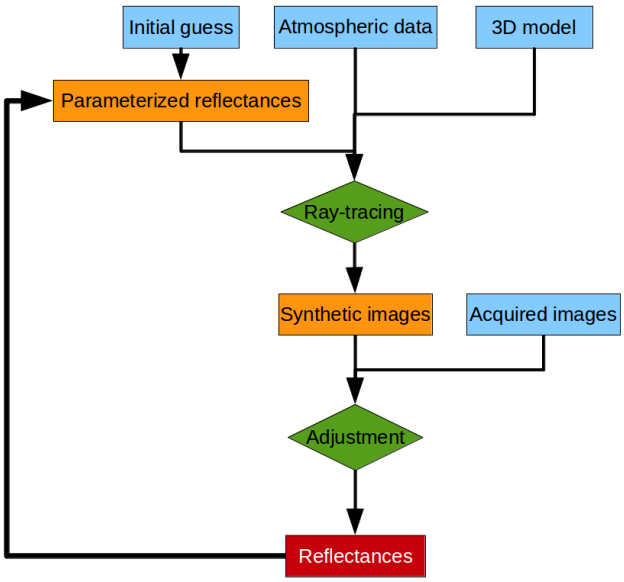


Figure 2: Principle of reflectance inversion.

4 INVERSE PROBLEM: REFLECTANCE ESTIMATION

4.1 General principle and plain algorithm

We face an inverse problem which direct problem has been discussed in the previous section. We consider that the unknown properties are the reflectances of the materials that compose the scene. That means that lighting conditions are supposed to be known, as well as the geometric 3D model (including the geometric calibration of the camera). The estimation of the reflectance from the images is a complex inverse problem in the general case (non-Lambertian materials, any geometry), because the total irradiance at a point P depends on the reflectance of the other surface elements of the scene, through the reflected irradiance. What is more, the directional effects of reflectance have to be modelized by a parametric function, the BRDF, that is not linear. So we use a minimization algorithm to estimate the parameters of the BRDF function of every material, in an iterative way, as shown on figure 2. The function to be minimized is the difference between

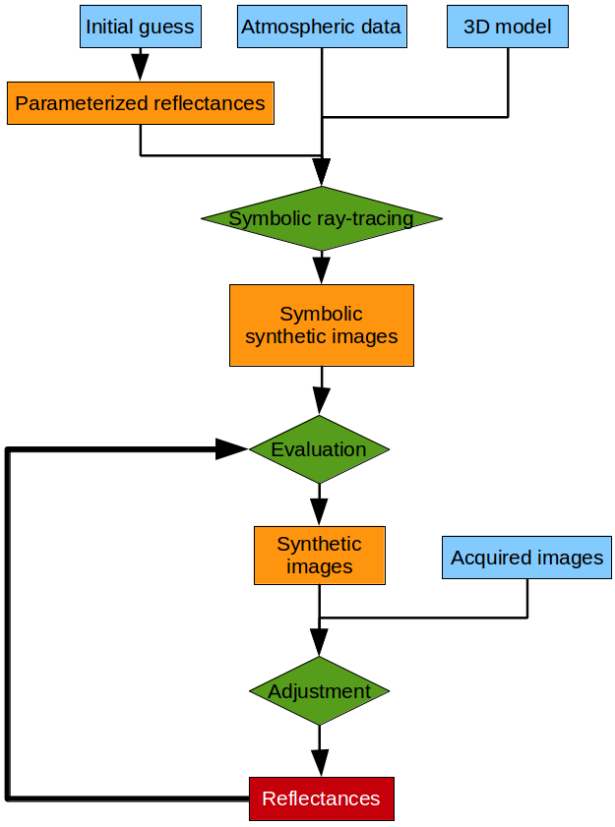


Figure 3: Inversion algorithm with symbolic ray-tracing.

acquired and synthetic sensor images:

$$\epsilon = \left(\sum_{image\ j} \sum_{pixel\ i} [L_{sensor}^{synthetic}(i,j,\boldsymbol{\rho}) - L_{sensor}^{acquired}(i,j)]^p \right)^{\frac{1}{p}} \quad (6)$$

This difference is the cost function of the set of parameters $\boldsymbol{\rho} = (\rho_i)_i$ of the BRDF models (*e.g.* albedos of the Lambertian materials). The input data is composed of the atmospheric properties (defining lighting conditions), the segmented geometric 3D model and initial reflectance guesses. They may be based on prior knowledge of the scene materials or on a preliminary inversion under a Lambertian hypothesis.

4.2 Symbolic ray-tracing for inversion

A major difficulty appears while minimizing (6): it requires the evaluation of the radiance image with different sets of parameters, naively leading to numerous computationally intensive ray-tracing renderings. For instance, for a gradient descent optimization relying on finite difference estimates, images not only have to be rendered for the current parameter values, but also for each dimension of the parameter vector $\boldsymbol{\rho}$. We can however observe that for a given camera position, scene geometry and illumination, the radiance can be expressed as a function of the BRDF parameters. Thus, by keeping the rendered radiance image as a function of the BRDFs parameters instead of a numerical value, the costly ray-tracing rendering may only be performed once. That is what we refer to as **symbolic ray-tracing**. For instance, if we assume that the materials of the scene are Lambertian, their BRDFs are each described by a single parameter: the albedos ρ_j, and $\dim(\boldsymbol{\rho})$ counts these Lambertian materials. Thanks to equations (1) and (2), the radiance carried by any light path C is of the form:

$$L(\boldsymbol{\rho}, C) = \alpha \prod_{j=1}^{\dim(\boldsymbol{\rho})} \rho_j^{e_j} \quad , \quad \text{with} \quad \begin{cases} \alpha \in \mathbb{R}, \quad e_j \in \mathbb{N} \\ \sum_j e_j \leq N_{max} \end{cases} \quad (7)$$

where α is a coefficient depending on the geometry, the atmospheric transmission and the power of the light source where the path originates. The exponent e_j is equal to the number of reflections of the path on material j (and thus 0 if that material is not hit by the path C). The radiance of a pixel being the Monte-Carlo sum (5) of the weighted contributions of many light paths, it may be expressed as a multivariate polynomial function:

$$L_{sensor}^{synthetic}(\boldsymbol{\rho}) = \sum_k \beta_k \cdot L(\boldsymbol{\rho}, C_k) = \sum_k \gamma_k \prod_{j=1}^{\dim(\boldsymbol{\rho})} \rho_j^{e_{j,k}} \quad (8)$$

It is easy to show that if the scene reflectances are all kernel BRDFs and not only Lambertian, the pixel radiance $L_{sensor}^{synthetic}$ is still a multivariate polynomial (8), where $\dim(\boldsymbol{\rho})$ is now the total number of BRDF kernel terms f_i (3).

Once the symbolic image has been computed, it is easier to minimize (6) with a standard minimization algorithm: the cost function and its derivative can be evaluated directly and quickly for any set of parameters. The inversion algorithm only needs a single ray-tracing rendering, as shown on figure 3. Furthermore, the minimization process avoids the instability induced by the change in the stochastic noise that would be implied by the several rendering passes of the plain algorithm (figure 2). Besides drastically speeding up the optimization process, symbolic ray-tracing thus improves its numerical stability.

4.3 Symbolic ray-tracing tractability

The number N_e of $(e_j)_j$ series satisfying constraints (7) is the number of N_{max} combinations with repetitions of $\dim(\boldsymbol{\rho}) + 1$ elements, known as the multiset coefficient $\binom{\dim(\boldsymbol{\rho})+N_{max}}{N_{max}}$. The space complexity of storing the multivariate polynomial function $L_{sensor}^{synthetic}(\boldsymbol{\rho})$ (8) is thus bounded by N_e since each of its monomial term corresponds to a unique series $(e_j)_j$. A monomial weight is thus simply a sum of γ_k values corresponding to identical series $(e_{j,k})_j$. For instance, for a scene with 10 different materials and at most 3 reflections, the maximal number of monomial terms in $L_{sensor}^{synthetic}(\boldsymbol{\rho})$ is $N_e = 286$.

In the general case of non-kernel BRDFs, this bounded size refactoring is however not possible, and thus the space-complexity of $L_{sensor}^{synthetic}(\boldsymbol{\rho})$ grows with each new sampled path of the Monte Carlo estimate (5). For instance the models from (Torrance and Sparrow, 1967), (Cook and Torrance, 1981) or (Ward, 1992), whose exponential term, containing both a geometric coefficient and a model parameter, prevents the factorization of the $L(C_k)$ terms in (5). As Monte-Carlo integration relies on many rays to decrease the stochastic noise; using non factorizable BRDF models becomes untractable, which drove our restriction to kernel BRDFs.

4.4 Implementation details

We implemented the symbolic ray-tracing in LuxRender, by keeping the expression of the reflectance every time a ray hits the surface of an object. The symbolic expressions are handled through the GPL library GiNaC. For minimizing (6), we chose a conjugate gradient method, using the GPL scientific library GSL. Furthermore, we chose $p = 2$ for the norm of the difference (6), because it is the lowest degree such that ϵ^p is a multivariate polynomial (hence without absolute differences and differentiable).

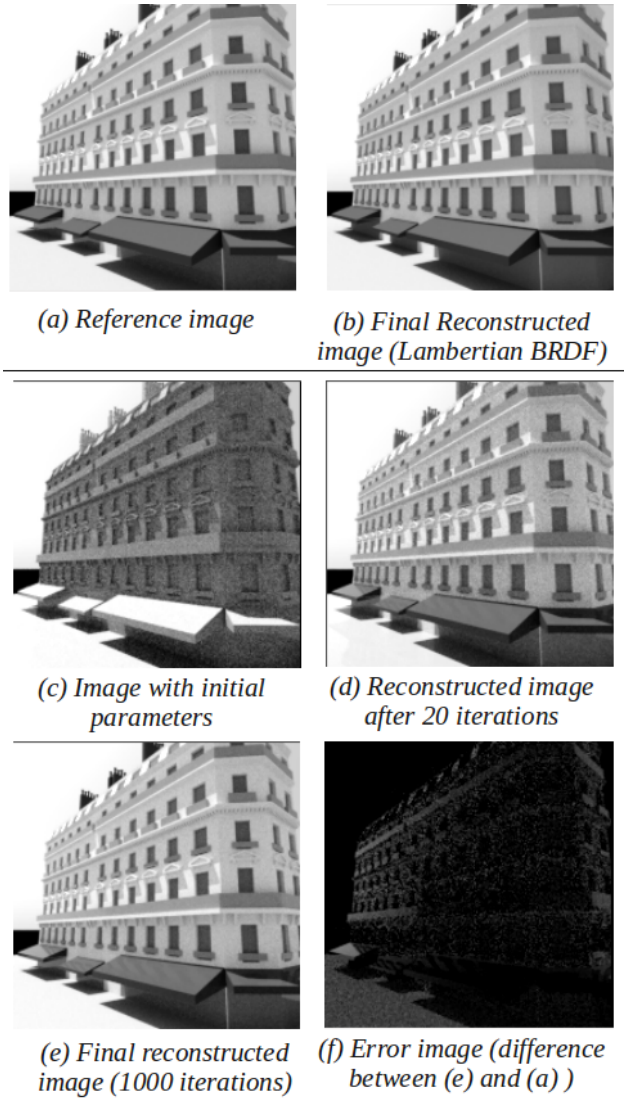


Figure 4: Reference image (a), reconstruction with Lambertian BRDF (b), and convergence of the estimation with kernel BRDF.

5 RESULTS OF THE INVERSION METHOD WITH SYNTHETIC DATA

5.1 Scene description and methodology

The inversion algorithm is tested on a synthetic urban scene (figure 4), at wavelength $\lambda = 550nm$. The geometry is described by a very detailed 3D model, and the material have a Lambertian BRDF or a BRDF with Lambertian and specular kernels (4) (for walls and windows). The environment map describing atmospheric scattering is computed with 6S. A reference image of the synthetic scene is simulated with LuxRender. A second simulation is made with symbolic ray-tracing, creating a symbolic image from the same point of view. The numerical and symbolic images cannot be simulated with the same ray-tracing pass, because it would lead to correlation between the stochastic noise due to the Monte-Carlo integration; that could enhance the inversion results without qualifying the method itself. Furthermore, the symbolic ray-tracing uses the numerical values of the BRDF parameters to drive importance sampling, so these values must be set arbitrarily.

Object	Reference albedo	Estimated albedo	Error
Roof	0.2	0.285	0.085
Wall	0.8	0.761	-0.039
Balcony	0.3	0.291	-0.009
Asphalt	0.2	0.201	0.001
Awning	0.05	0.0496	-0.0004
Window	0.9	0.230	-0.670

Table 1: Inversion results on a synthetic urban scene, with a single image, assuming a Lambertian BRDF model

Object	Param.	Ref. value	Estim. value	Error
Roof	ρ_d	0.2	0.390	0.190
Wall	ρ_d	0.72	0.762	0.042
	ρ_s	0.08	0.326	0.246
Balcony	ρ_d	0.3	0.284	-0.016
	ρ_s	0	0.373	0.373
Asphalt	ρ_d	0.2	0.218	0.018
	ρ_s	0	0.000	0.000
Awning	ρ_d	0.05	0.046	-0.004
	ρ_s	0	0.300	0.300
Window	ρ_d	0.18	0.185	0.005
	ρ_s	0.72	0.494	-0.226

Table 2: Inversion results on a synthetic urban scene, with a single image, assuming a Lambertian+Specular kernel BRDF model

5.2 Parameter estimation

Lambertian BRDF. We first perform the symbolic simulation assuming that the materials have a Lambertian BRDF. The diffuse part of the reference BRDFs are correctly estimated (table 1), except for materials that are not directly seen in the image (such as roofs), as well as for materials with a high specular component (such as windows).

Lambertian+Specular BRDF. In order to retrieve the specular behaviour of materials, we make a symbolic simulation assuming that all materials have a Lambertian+specular kernel BRDF (4). The results of the parameter estimation are shown in table 2. The total process (symbolic ray-tracing and minimization) takes about 20 min, with images of 200x200 pixels. Though the value itself of the parameters can be estimated with high errors, the image reconstructed with these parameters is visually close to the reference image (figure 4). This is due to the ill-posedness of the inverse problem when trying to estimate the specular component, which is not always directly seen by the camera for each material. For instance there is no point light reflection visible in the windows, but only a large surface light (the sky). There is therefore a strong correlation between the effects of the specular and diffuse components.

6 CONCLUSION AND FUTURE WORK

In this work, we have developed an inversion algorithm to estimate the reflectance of materials of an outdoor scene, assuming known atmospheric conditions, and a known and segmented geometric 3D model. The geometric complexity of the 3D model only impacts the computing time of the symbolic ray tracing during the preprocessing step. The inversion method uses symbolic ray-tracing, therefore it can perform reflectance estimation using a single rendering pass. It proves to yield good results with synthetic data for Lambertian BRDF models, but has difficulties for

estimating the specular components of Lambertian+specular kernel BRDFs. The estimated reflectances of the materials can be used to reconstruct images of the scene from any point of view.

To improve the results, we could use the redundant information of multi-view acquisition (that is the case for terrestrial acquisitions with a vehicle). It cannot be done for complex scenes without improving our current code, because it is highly memory-consuming. However, as the image reconstructed from the estimated parameters with Lambertian models is close to the reference image, the algorithm should be tested on a real set of urban terrestrial images, to qualify the reliability of the Lambertian hypothesis for relighting purpose.

From a more theoretical point of view, the space-complexity of the symbolic ray-tracing can be limited by using of a decomposition of BRDF in an adapted function basis, leading to a kernel BRDF model; see (Rusinkiewicz, 1997) for a review on BRDF decomposition. That decomposition can better represent the physical behaviour of materials, that includes non-delta specular and retro-specular lobes. But the high number of parameters introduced by the decomposition leads to an ill-posed problem, and cannot be used in our context.

Another strong hypothesis used in this work is the segmentation of the 3D model into homogeneous regions; textured segments may thus not be processed. For instance, (Yu et al., 1999) consider a spatially-varying diffuse albedo, while the specular part is constant per area. This can lead to a better modelization of the materials of a real scene, which are not homogeneous, but increases the number of parameters to be estimated.

ACKNOWLEDGMENTS

The authors would like to thank Thales Training and Simulation for providing the segmented 3D façade model used for inversion which was produced during the Terra Numerica project of the Cap Digital cluster based on street-level data acquired by a mobile mapping system.

References

Boivin, S. and Gagalowicz, A., 2001. Image-based rendering of diffuse, specular and glossy surfaces from a single image. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pp. 107–116.

Chandelier, L. and Martinoty, G., 2009. A radiometric aerial triangulation for the equalization of digital aerial images and orthoimages. Photogrammetric Engineering and Remote Sensing.

Cook, R. and Torrance, K., 1981. A reflectance model for computer graphics. In: Proceedings of the 8th annual conference on Computer graphics and interactive techniques, ACM, pp. 307–316.

Devaux, A. and Paparoditis, N., 2010. Increasing Interactivity in Street View Web Navigation Systems. ACM Multimedia.

Dutre, P., Bala, K. and Bekaert, P., 2006. Advanced global illumination. AK Peters Ltd.

Kajiya, J., 1986. The rendering equation. In: Proceedings of the 13th annual conference on Computer graphics and interactive techniques, ACM, pp. 143–150.

Lachérade, S., Miesch, C., Boldo, D., Briottet, X., Valorge, C. and Le Men, H., 2008. ICARE: A physically-based model to correct atmospheric and geometric effects from high spatial and spectral remote sensing images over 3D urban areas. Meteorology and Atmospheric Physics 102(3), pp. 209–222.

Lensch, H., Kautz, J., Goesele, M., Heidrich, W. and Seidel, H., 2003. Image-based reconstruction of spatial appearance and geometric detail. ACM Transactions on Graphics (TOG) 22(2), pp. 234–257.

Machida, T., Takemura, H. and Yokoya, N., 2007. Inverse reflectometry for real objects with diffuse and specular interreflections. Electronics and Communications in Japan (Part II: Electronics) 90(1), pp. 50–60.

Martinoty, G., 2005. Reconnaissance de matériaux sur des images aériennes en multirecouvrement, par identification de fonctions de réflectances bidirectionnelles. PhD thesis, Université Paris 7.

Meyzonnette, J. and Lépine, T., 1999. Bases de radiométrie optique. Cépaduès.

Miesch, C., Briottet, X., Kerr, Y. and Cabot, F., 2000. Radiative transfer solution for rugged and heterogeneous scene observations. Applied Optics 39(36), pp. 6830–6846.

Müller, P., Zeng, G., Wonka, P. and Van Gool, L., 2007. Image-based procedural modeling of facades. ACM Transactions on Graphics 26(3), pp. 85.

Nicodemus, F., Richmond, J., Hsia, J. and Ginsberg, I., 1977. Geometrical considerations and nomenclature for reflectance.

Patow, G. and Pueyo, X., 2003. A survey of inverse rendering problems. In: Computer graphics forum, Vol. 22number 4, pp. 663–687.

Pharr, M. and Humphreys, G., 2004. Physically based rendering: from theory to implementation. Morgan Kaufmann.

Rusinkiewicz, S., 1997. A Survey of BRDF Representation for Computer Graphics. Technical report, Princeton University.

Torrance, K. and Sparrow, E., 1967. Theory for off-specular reflection from roughened surfaces. Journal of the Optical society of America 57(9), pp. 1105–1114.

Veach, E., 1997. Robust Monte Carlo methods for light transport simulation. PhD thesis, Stanford University.

Vermote, E., Tanré, D., Deuzé, J., Herman, M. and Morcette, J., 2002. Second simulation of the satellite signal in the solar spectrum, 6S: an overview. Geoscience and Remote Sensing, IEEE Transactions on 35(3), pp. 675–686.

Ward, G., 1992. Measuring and modeling anisotropic reflection. In: Proceedings of the 19th annual conference on Computer graphics and interactive techniques, ACM, pp. 265–272.

Yu, Y., Debevec, P., Malik, J. and Hawkins, T., 1999. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pp. 215–224.

FAST AND ACCURATE VISIBILITY COMPUTATION IN URBAN SCENES

Bruno Vallet and Erwann Houzay

Université Paris Est, IGN, Laboratoire MATIS
4, Av. Pasteur
94165 Saint Mandé Cedex, FRANCE
bruno.vallet@ign.fr (corresponding author), erwann.houzay@ign.fr
http://recherche.ign.fr/labos/matis

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: Photogrammetry, Visibility, Mobile mapping, Building facade, Urban Scene

ABSTRACT:

This paper presents a method to efficiently compute the visibility of a 3D model seen from a series of calibrated georeferenced images. This is required to make methods aimed at enriching 3D models (finer reconstruction, texture mapping) based on such images scalable. The method consists in rasterizing the scene on the GPU from the point of views of the images with a known mapping between colors and scene elements. This pixel based visibility information is then made higher level by defining a set of geometric rules to select which images should be used when focusing on a given scene element. The visibility relation is finally used to build a visibility graph allowing for efficient sequential processing of the whole scene

1 INTRODUCTION

One of the most investigated applications of mobile mapping is the enrichment an existing 3D urban model (obtained from aerial photogrammetry for instance) with data acquired at the street level. In particular, the topics of facade reconstruction (Frueh and Zakhor, 2003) and/or texture mapping (Bénitez et al., 2010) have recently received a strong interest from the photogrammetry and computer vision communities. An important step in such applications is to compute the visibility of the scene, that is to answer to the questions:

- Which scene element is visible at a given pixel of a given image (pixel level)
- Which scene elements are seen well enough in a given image, i.e. for which scene elements does the image contain significant information (image level)

Depending on the applications, these questions can be asked the other way, that is querying the images/pixels seeing a given scene element or 3D point. In practice, mobile mapping generates a large amount of data, and an acquisition led on an average sized city will generate tens of thousands of images of tens of thousands of scene elements. In that case, a per pixel visibility computation based on ray tracing as done usually (Bénitez and Baillard, 2009) becomes prohibitively costly.

The first part of this paper proposes instead a visibility computation method taking advantage of the graphics hardware to make this computation tractable. Another problem arising in handling a large number of images acquired in an urban environment is that a given scene element might be seen at various distances and with various viewing angles during the acquisition. As a result, the set of images that view a given scene element will be heterogeneous, thus hard to handle in most applications. The second part of this paper tackles this issue by proposing appropriate geometric criteria in order to reduce this set to a more homogeneous one, and where the images contain substantial information about the scene element.

The contributions of this paper are twofold, and both aim at facilitating the exploitation of mobile imagery at large scale in urban areas:

- Fast per pixel visibility complex computation (Section 2)
- Quality aware and memory optimized image selection (Section 3)

Results are presented and discussed in Section 4, and conclusions and future works are detailed in Section 5.

1.1 Previous works

Enriching an existing 3D city model with ground based imagery is a quite important topic in photogrammetry (Haala, 2004), as well as a major application of mobile mapping systems (Bénitez et al., 2010). More precisely, the visibility computation method presented in this paper is designed to allow scalability of:

- Facade texture mapping methods: the images are used to apply a detailed texture to the facades of the model. This is especially important for 3D models built and textured from aerial images for which the facades textures are very distorted.
- Reconstruction from images: the images are used to build a detailed 3D model of objects of interest such as facades (Pénard et al., 2004) or trees (Huang, 2008). In this case, a careful selection of the images to input the reconstruction method should be made. This paper proposes an efficient approach to automatize this selection as long as a rough 3D model of the object to reconstruct is known (rectangle for a facade, sphere or cylinder for a tree), which is required to make such methods applicable to large scale reconstructions.

In most reconstruction/texture mapping work relying on mobile imagery, the problem of visibility computation is raised but often not explicitly tackled, so we guess that the image selection

is either made manually from visual inspection, probably using a GIS. Whereas this is sufficient to demonstrate the pertinence of a method, this becomes a major lock when considering scalability to an acquisition led over an entire city. In order to solve this problem, (Bénitez and Baillard, 2009) proposes and compares three approaches to visibility computation: 2D/3D ray tracing and Z buffering. The 2D approach is the fastest but the 2D approximation fails to give the correct visibility in all the cases when buildings are visible behind others, which frequently occurs in real situations. 3D ray tracing and Z buffering do not have this limitation but computation time is important even for a very sparse sampling of the solid angle of the image. This can be an issue as if too few rays are traced, scene elements seen by a camera may be missed. Moreover, this is insufficient to handle occlusions at the scale of a pixel.

As mentioned in (Bénitez et al., 2010), facade texture mapping calls for a per pixel visibility computation to take into account two kinds of occlusions: *predictable* occlusions that may be predicted by the model to texture (parts of the model hiding the facade to texture in the current view), and *unpredictable* occlusions. Unpredictable occlusions can be tackled based on laser and/or image (Korah and Rasmussen, 2008) information. Predictable occlusions require to compute visibility masks of the 3D objects to texture (Wang et al., 2002). Such per pixel visibility computation is extremely costly, and the purpose of this paper is to make it tractable. We achieve this by exploiting the GPU rendering pipeline, which is known to be much faster than the CPU as it is extremely optimized for that purpose, and is in fact much more adapted to our problem.

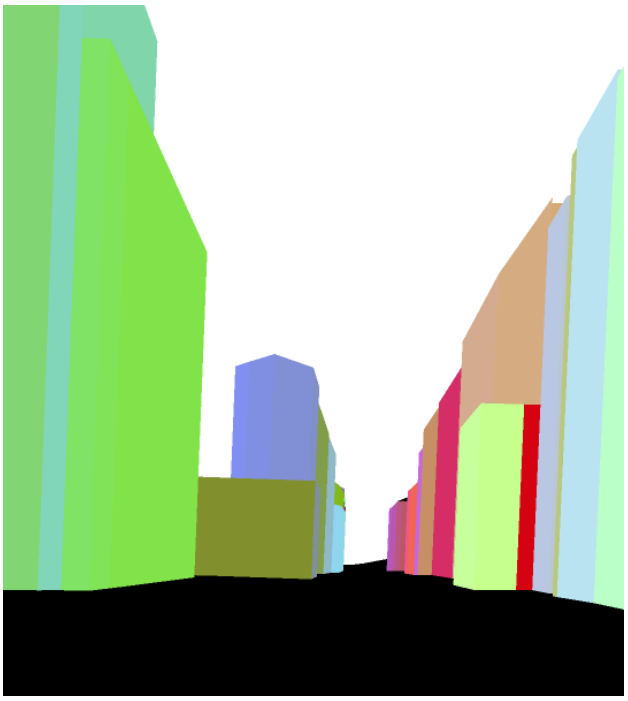
It should be mentioned that an alternative to visibility computation exists for 3D model enriching from ground data. It consists in constructing a textured 3D model from the ground data alone, then registering this ground based model to the 3D model to enrich. This is the preferred methodology in case laser scanning data was acquired simultaneously to the images. For instance, in (Frueh and Zakhor, 2003) and (Frueh and Zakhor, 2004) a ground based model is registered to the initial 3D city model based on Monte Carlo localization, then both models are merged. More recently, this was done based on images only with a reconstruction obtained by SLAM and registered by an Iterative Closest Point (ICP) method (Lothe et al., 2009). This could also be applied to the purely street side city modeling approach of (Xiao et al., 2009).

2 PIXELWISE VISIBILITY

This section presents the core of our method, which is to efficiently compute the visibility complex pixelwise by rasterization on the GPU. More precisely, we aim at finding which object of the 3D scene should be visible on each pixel of each acquired image. The obvious way to do that is to intersect the 3D ray corresponding to that pixel with the 3D scene, which is very costly even when optimizing this computation based on spatial data structures. In order to reduce computation time, we propose instead to rasterize the 3D scene from the viewpoint of each camera, that is to simulate the real life acquisition using the highly optimized rendering pipeline of the GPU. It relies on two successive steps:

1. Build a virtual camera in the 3D model accordingly to the extrinsic and intrinsic parameters of our real life camera (Section 2.1)
2. Perform a rasterization of the scene (see Fig. 1(b)) in a GPU buffer with a unique identifier per object of interest (Section 2.2).

(a) Real image



(b) Virtual image

Figure 1: A real image and superposable virtual image obtained by capturing the 3D model from the same point of view

2.1 Camera parameters transfer

Extrinsic parameters: Building an OpenGL camera simply requires to set the appropriate 4 by 4 matrix corresponding to the projective transform. However, special care should be taken as GPUs usually work in single precision floating point coordinates, as their development has always been driven by the games industry that focuses more on performance than on precision. This is completely incompatible with the use of geographic coordinates, and can lead to unacceptable imprecision during the rendering. A simple workaround is to choose a local frame centered within

the 3D scene, in which both the 3D model coordinates and the camera orientation will be expressed.

Intrinsic parameters: Calibration of a camera sets its intrinsic parameters consisting mainly in its focal, principal point of autocollimation (PPA) and inner distortion. Conversely, OpenGL relies on the notion of field of view (FOV), the PPA is always at the perfect center of the image and distortion cannot be applied. In most cases, the PPA is close enough to the image center for this to be negligible. If it is not, a larger image should be created containing the image to be rendered and with its center at the PPA, then cropped. Finally, the field of view should then be be computed by:

$$FOV = 2. * \tan^{-1}\left(\frac{\max(width, height)}{2f}\right); \qquad (1)$$

A simple means to handle the distortion is to resample the rendered image to make it finally perfectly superposable to the acquired one. Depending on the application, this resampling might not be necessary, and it might be more efficient to apply the distortion on the fly when a visibility information is queried. In both cases, the width and height of the rendered image should be chosen such that its distortion completely encompasses the acquired image.

2.2 Rasterization

The problem of *rendering* (generating virtual images of 3D scenes) has been widely studied in the computer graphics community, and two main methodologies have arisen:

1. **Rasterization** consists in drawing each geometric primitive of the scene using a depth buffer (also called Z-buffer) to know which primitive is in front of which from the current viewpoint.

2. **Ray tracing** consist in intersecting each 3D ray corresponding to a screen pixel with the scene, then iterating on reflected rays in order to define the appropriate color for that pixel.

Ray tracing is known to be much more expensive, but allows for very realistic lighting and reflexion effects. Ray tracing is also much harder to parallelize, such that GPUs always perform rendering by rasterization, making it extremely efficient and well suited in our case.

The result of a rasterization is a color image (the Z-buffer is also accessible if needed). Thus, the simplest way to get a visibility information from a rasterization is to give a unique color to each scene object of interest, and create a mapping between colors and scene objects. The only limitation to this method is that the number of objects of interests should not exceed the number of colors ($256^3 = 2^{24} \approx 16.8$ million) which in practice is completely sufficient (there are much less objects of interest such as facades or trees in the largest cities).

The second problem is that the size of the image to render might be arbitrarily larger than the size of the screen of the computer on which we will run the visibility computation. Hopefully, GPUs can perform *offline rendering*, that is to render a scene in a buffer of the GPU which size is only limited by the graphical memory (GRAM). This buffer can then be transferred to the RAM, and saved on the disk if required. Using lossless compression is strongly advised as colors now have an exact meaning that should absolutely be preserved, and most images rendered this way only present a limited number of colors arranged in large objects allowing for high lossless compression rates. In our experiments, the 1920x1080 buffers stored this way weighted 12 KBytes in average, which corresponds to a compression rate of 99.8%.

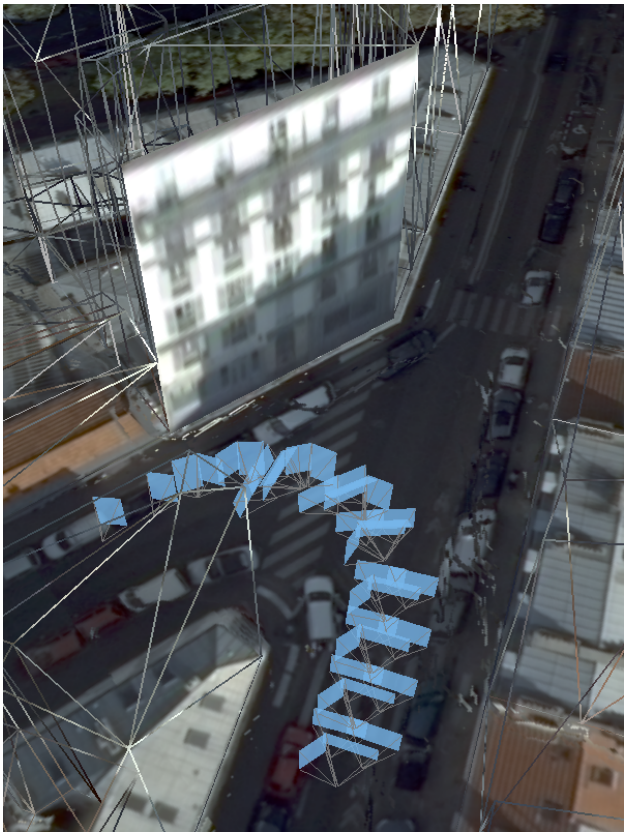


Figure 2: Set of images selected by our method as containing pertinent information about a facade (non wireframe)

The result of this step, that we will call *visibility image* (Fig. 1), stores the direct visibility information, i.e. the answer to the question: Which is the object seen at a given pixel of a given image? Some more work still needs to be done in order to answer efficiently to the image level question: which objects are seen well enough in a given image ? And the inverse one: which images see a given object well enough? The next section proposes a method to answer this questions that relies on a proper definition of the term "well enough" in this context.

3 PICTURE LEVEL VISIBILITY

Most reconstruction/texture mapping problems can be decomposed into one sub-problem for each scene element of interest. In this case, we need to be able to know which images will be useful to process the scene element in order to load them into the computer's memory. Loading too many will impair processing time and memory footprint, but it can even lower the quality of the result in some cases. This section tackles these problems in two steps:

1. Defining geometric criteria to select which images are useful for the processing.

2. Building a visibility graph to answer to inverse visibility queries.

Optionally, we will propose two applications of the visibility graph structure:

1. Optimizing the memory handling in sequential treatment.

2. Selecting a single sequence to avoid limit issues due to non static scene elements (shadows, mobile objects,...).

3.1 Geometric criteria

The first naive approach to determining the scene elements seen in an image is simply to build a list of all colors present in a visibility image, and declare the corresponding scene elements as viewed by the image. This is both inefficient and too simplistic. In fact, we want to select the images to use to process a scene element, so only the images seeing that element well enough should be used. Thus we propose two geometric criteria to select only the appropriate images:

1. **Image content**: The image should contain sufficient information about the scene element, or in other terms, the scene element should cover at least a certain portion of the visibility image. This criterion also has a practical aspect: we can accelerate the color counting by only checking the colors of a small sparse subset of the image. For instance, if the criterion corresponds to 10 000 pixels, checking only every 1 000 pixel is sufficient, provided that the sampling is homogeneous enough, and the criterion becomes having at least 10 samples of a given color. This will accelerate the pixel counting without impairing the result.

2. **Resolution**: The size of the pixel projected on the scene element should not exceed a given value (say $25cm^2$) or a given multiple (say 5 times) of the smallest projected pixel size. We estimate this projected pixel size at the barycenter of the projection of the scene element in the visibility image (this can be done simultaneously to the color counting, using the same sparse sampling). This criterion penalizes both distance and bad viewing angles.

These criteria should be sufficient for most applications, but they can be adapted if needed. An example of the set of images selected for a given scene element (facade) based on these criteria is shown on Fig. 2.

3.2 The visibility graph

The geometric criteria cited above establish a relation: image I_i sees element E_j well enough, or conversely element E_j is seen in image I_i well enough. This relation is built from the image point of view as for each image we define which elements are seen well enough. However, we usually need the inverse information: for the scene element on which we want to focus our method, which image should we use ? This requires to build a visibility graph containing two types of nodes (image and scene elements). Each visibility relation will be an edge in this graph between an image and an element node, that will be registered from both image and edge point of view. Constructing this graph is required to inverse the visibility information, but it is also useful for optimization and optionally for further simplification. A visualization of this visibility graph is proposed in Fig. 3.

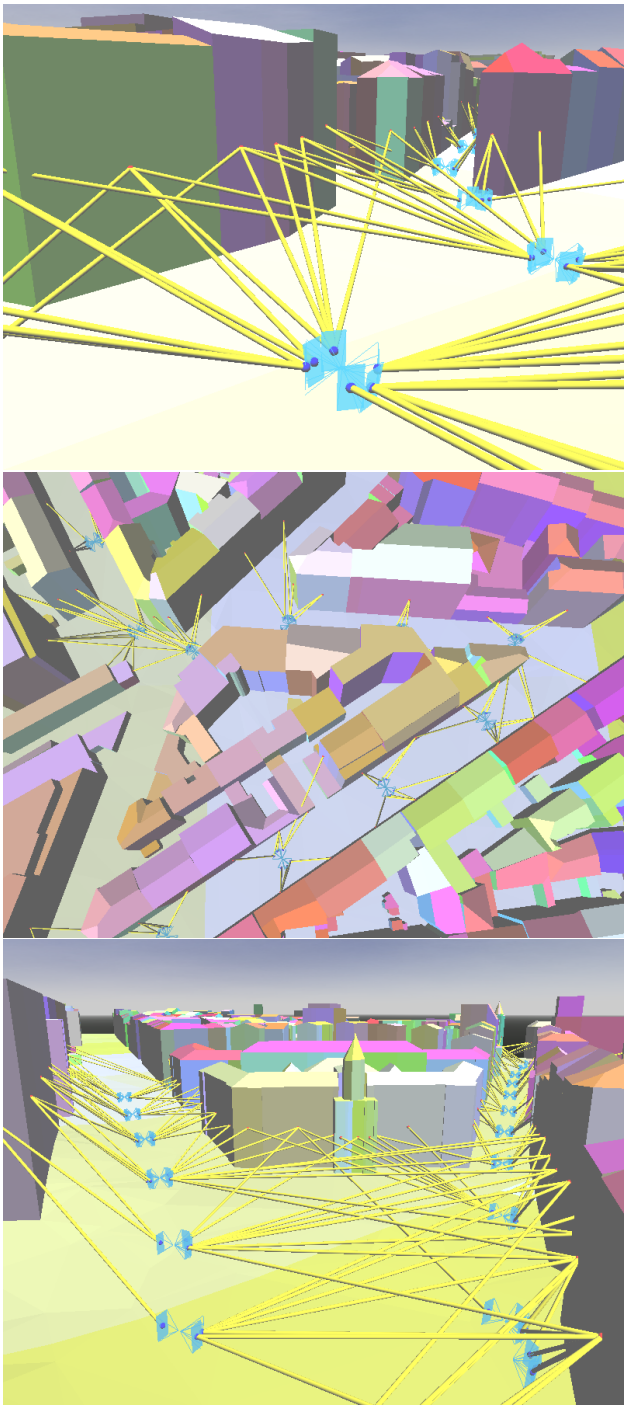


Figure 3: Various views of the visibility graph computed on our test scene (for clarity, only 10% of the images were used). Image (resp. scene) nodes are displayed in blue (resp. red).

3.3 Optimizing memory handling

The visibility graph allows to cut a reconstruction/texture mapping problem into sub-problems by inputing only the images seeing a scene element well enough. A trivial approach to process an entire scene sequentially is to load all these images when processing each scene element, then free the memory before processing the next element. This trivial approach is optimal in terms of memory footprint, but an image seeing N scene elements will be loaded N times, which increases the overall computing time, especially if the images are stored on a Network Attached Storage (NAS). Using a NAS is common practice in mobile mapping as

Dataset	#walls	#images	path length
Ours	4982	5344 (334x16)	1km
Bénitez	11408	1980 (990x2)	4.9km

Table 1: Comparison between our dataset and the dataset of Bénitez.

an acquisition produces around one TeraByte of (uncompressed) image data per hour. The optimization we propose is useful if the loading time is not negligible compared to the processing time, and if the network is a critical resource (in case images are on a NAS). It consists in finding the appropriate order in which to process the scene elements in order to load each image only once. The algorithm is based on the visibility graph, where additional information will be stored in each node (loaded or not for images, processed or not for scene elements):

1. Select any scene element E_j in the scene as starting point of the algorithm and add it to a set S_a of "active elements".
2. For each image I_i^j viewing E_j, load it and add the unprocessed elements seen by I_i^j to S_a.
3. Process E_j, mark E_j as processed and remove it from S_a.
4. If an I_i^j has all its viewed elements processed, close it.
5. Select the element E_j with fewest unopened seeing images in S_a.
6. While S_a is not empty, go back to 2.
7. If no unprocessed element remain, terminate.
8. Select an unprocessed element E_j and go back to 2,

This algorithm is quite simple to implement once the visibility graph has been created, and will be evaluated in Section 4.

3.4 Sequence selection

Another useful utilization of the visibility graph is *sequence selection*. In practice, we found out that images acquired by a mobile acquisition device are often redundant, as the coverage of an entire city require to traverse some streets more than once. Most georeferencing devices use an inertial central allowing for very precise relative localization but can derive due to GPS masks. Hence redundant image sequences viewing the same scene element usually have a poor relative localization. Moreover, the scene may have changed between two traversals: parked cars gone, windows closed, shadows moved... In consequence, we propose to cluster the set of images seeing a given scene element according to their time of acquisition, then select the cluster (sequence) with the best quality (using the criteria of Section 3.1).

4 RESULTS AND DISCUSSION

We have developed the tools described in this paper in order to perform large scale reconstruction and texture mapping of various urban objects such as facades, trees,... However, this paper only focuses on the optimized visibility computation, that is a mandatory prerequisite for such applications. Consequently, the results presented in this section consist mainly in statistics and timings demonstrating the quality efficiency of our approach.

The method presented in this paper (rasterization) was evaluated on a set of images acquired in a dense urban area with the mobile mapping system of (Bentrah et al., 2004). The set consists

Figure 4: Visualization of the 5344 images used in our experiments.

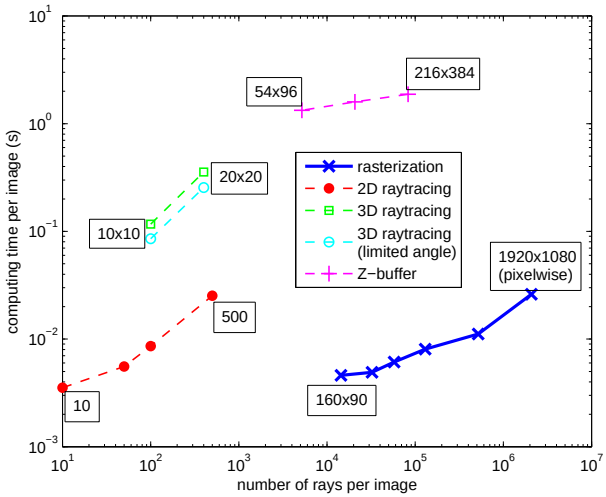


Figure 5: Timings for visibility computation. Results are compared between our rasterization approach (thick blue line) with subsampling factors ranging from 1 to 24, and the three approaches of Bénitez (dotted lines).

of 334 vehicle positions, every 3 meters along a 1km path. For each position, 16 images were acquired (12 images forming a panoramic + 2 stereo pairs). We also disposed of a 3D city model of the acquired area built from aerial imagery. Comparison with the dataset of (Bénitez and Baillard, 2009) is displayed in Table 1. The main difference is that our path is shorter, but our image density is much higher. The images acquired are represented inserted in the 3D model on Fig. 4.

The timings for the computation of the visibility graph running on an NVidia GeForce GTX 480 are presented on Fig. 5, along with equivalent timings taken from (Bénitez and Baillard, 2009). They are given with respect to the number of visibility queries computed per image (number of rays traced or of buffer pixels). A high number of rays ensures that most visibility relationships will be found (good angular precision).

As expected, the use of the GPU allows for a huge performance increase for 3D visibility computation, even though our results are much more accurate: for an equivalent number of rays, our rasterization approach is around 400 times faster than Z-buffering. This huge performance increase is however limited to high ray numbers, as decreasing the resolution of our rasterization does

min area	max pix size	image/ wall	mode	max footprint	#image loads
0	∞	39.2	trv.	414	52085
			opt.	1854	5058
1%	$(10cm)^2$	28.7	trv.	191	9644
			opt.	521	3811
2%	$(4cm)^2$	20.1	trv.	120	5337
			opt.	242	2886
5%	$(2cm)^2$	13.9	trv.	44	2665
			opt.	51	1793

Table 2: Influence of geometric criteria on visibility graph density, and comparison of memory footprint and loading times between trivial (trv.) and optimized (opt.) processing.

not improve the computing time below 10^4 rays. Some steps of the rendering pipeline have a computation time depending on the number of primitives of the scene and not on the size of rendering, which explains the limit we reach (around 4.6ms per image) for low resolutions. Even though our computing time does not improve below 10^4 rays, it is still 20 times faster (for 160x90 rays) than 10x10 3D ray tracing.

The only method of (Bénitez and Baillard, 2009) with performance comparable to ours is 2D ray tracing. In our sense, this method is not suitable in urban areas in which configurations requiring the third dimension are often encountered (such as a taller building behind a smaller one). This is confirmed by (Bénitez and Baillard, 2009) who found that 2D ray tracing misses one third of the walls (compared to Z-buffering). Moreover, they state that 100 rays per image is a good compromise. For this number of rays, 2D ray tracing is already slower than our rasterization.

Finally, our approach is clearly the only one allowing for pixelwise visibility computation in reasonable time (26ms per image, rightmost point of Fig. 5). It can even be brought down to 20ms per image with the color counting acceleration described in Section 3.1. This high performance makes pixelwise visibility image computation time of the same order of magnitude than image loading time (15ms in our experiments), so the visibility images can be computed on the fly when required instead of being precomputed and saved, which is another nice performance improving feature of our approach.

Finally, we evaluated the optimized memory handling by creating four visibility graphs of different densities by imposing increasingly harsh geometric constraints (see Table 2). As expected, optimized processing greatly reduces the number of loads (each image is loaded exactly once) at the cost of memory footprint (maximum number of images loaded simultaneously), and this effect is more important on denser graphs. If memory size is a limit and/or processing time is large compared to data loading time, then the trivial approach should be used. In other cases, and especially if data transfers are the bottleneck, then the optimized method will be preferable. Table 2 also shows that harsher constraints reduces the number of selected images, such that only the most pertinent ones are preserved.

5 CONCLUSIONS AND FUTURE WORK

We have presented a methodology allowing easy scaling of reconstruction and texture mapping methods on large areas. Computing the visibility graph of large scenes becomes tractable based on our approach, even at the pixel level. Enriching 3D models based on large amounts of data acquired at the ground level is becoming a major application of mobile mapping, and we believe that this methodology will prove useful to make the algorithms developed in this context scalable. Our evaluation shows that our approach outperforms previous works both in quality and computing time.

The method described in this paper is mostly useful for texture mapping purposes where the per pixel visibility information is required in order to predict occlusion of the model by itself. However, it can also be used for unpredictable occlusions by inserting a point cloud corresponding to detected occluders in the 3D scene before rendering. But this method can also be used for any reconstruction method where a rough estimate of the geometry is known (bounding box, point cloud, 2D detection in aerial images,...)

In the future, we will look into optimizing our approach for larger 3D models based on spatial data structures such as octrees in order to load only the parts of the model that are likely to be seen. We will also investigate doing the color counting directly on the GPU to avoid transferring the buffer from graphics memory to RAM.

REFERENCES

Bénitez, S. and Baillard, C., 2009. Automated selection of terrestrial images from sequences for the texture mapping of 3d city models. In: CMRT09. IAPRS, Vol. XXXVIII, Part 3/W4, pp. 97–102.

Bénitez, S., Denis, E. and Baillard, C., 2010. Automatic production of occlusion-free rectified faade textures using vehicle-based imagery. In: IAPRS, Vol. XXXVIII, Part 3A (PCV'10).

Bentrah, O., Paparoditis, N. and Pierrot-Deseilligny, M., 2004. Stereopolis : An image based urban environments modelling system. In: International Symposium on Mobile Mapping Technology (MMT), Kunming, China, March 2004.

Frueh, C. and Zakhor, A., 2003. Constructing 3d city models by merging ground-based and airborne views. In: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR03).

Frueh, C. and Zakhor, A., 2004. An automated method for large-scale, ground-based city model acquisition. In: International Journal of Computer Vision, Vol. 60(1), pp. 5–24.

Haala, N., 2004. On the refinement of urban models by terrestrial data collection. Arch. of Photogrammetry and Remote Sensing, Commission III WG 7 35, pp. 564–569.

Huang, H., 2008. Terrestrial image based 3d extraction of urban unfoliaged trees of different branching types. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVII, Beijing, China.

Korah, T. and Rasmussen, C., 2008. Analysis of building textures for reconstructing partially occluded facades. In: European Conference on Computer Vision.

Lothe, P., Bourgeois, S., Dekeyser, F., Royer, E. and Dhome, M., 2009. Towards geographical referencing of monocular slam reconstruction using 3d city models: Application to real-time accurate vision-based localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami, Florida.

Pénard, L., Paparoditis, N. and Pierrot-Deseilligny, M., 2004. 3d building facade reconstruction under mesh form from multiple wide angle views. In: IAPRS vol. 36 (Part 5/W17), 2005.

Wang, X., Totaro, S., Taillandier, F., Hanson, A. R. and Teller, S., 2002. Recovering facade texture and microstructure from real-world images. In: Proc. of the 2nd International Workshop on Texture Analysis and Synthesis in conjunction with ECCV'02, pp. 145–149.

Xiao, J., Fang, T., Zhao, P., Lhuillier, M. and Quan, L., 2009. Image-based street-side city modeling. In: ACM Transaction on Graphics, 28(5), 2009 (also in proceedings of SIGGRAPH ASIA'09).

Part 2

Papers accepted by extended abstract review

QUALITY ASSESSMENT OF LANDMARK BASED POSITIONING USING STEREO CAMERAS

S. Hofmann, M. J. Schulze, M. Sester, C. Brenner

Institute of Cartography and Geoinformatics, Leibniz Universität Hannover, Appelstraße 9a, 30167 Hannover, Germany – (sabine.hofmann, maltejan.schulze, monika.sester, claus.brenner)@ikg.uni-hannover.de

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: Stereo Camera, Accuracy, Simulation, Mobile Mapping, Geometry, Navigation, Reference Data, Quality

ABSTRACT:

Driving autonomously requires highly accurate positioning. Therefore, alternative positioning systems to GPS are required especially to increase the accuracy, and to have a complementary data source in areas where GPS is not available. As more and more on-board sensors are used for safety reasons, information gathered about their environment can be used for positioning based on relative measurements to landmarks along the road. This paper investigates the accuracy potential of positioning using a stereo camera system and landmark maps. Therefore, we simulated several stereo camera systems with variable opening angle and base length to compute the positioning accuracy in a test area. In the first step, localization was calculated based on single positions, in the second step we used a Kalman filter additionally. While positioning in the first case was not successful along the entire trajectory, the Kalman filter led to far better results.

1. INTRODUCTION

Driving autonomously or driver assistance applications make highly accurate positioning much more important than in today's navigational devices. Therefore, alternative positioning systems to GPS are required to increase the accuracy, and to have a complementary data source in areas where GPS fails, for example in street canyons.

On-board sensors, such as cameras, laser scanner and radar, which gather information about their environment for active safety systems, for example environment detection to avoid collisions with pedestrians or recognition of traffic signs, can be used for positioning based on relative measurements to objects along the road. To use this data for positioning a highly accurate representation of the environment is required. Iconic representations, for example occupancy grids or symbolic representations, for example landmark based maps (Burgard and Hebert, 2008) can be used. Using landmarks such as poles of traffic signs and traffic lights for positioning has been investigated earlier, e.g. by Weiss et al., 2005, Brenner, 2010.

To investigate, how accurate positioning is possible using a stereo camera system and landmark maps, we simulated several stereo camera systems with variable opening angle and base length. In addition to calculating the accuracy of single positions, we simulated the knowledge of two different types of inertial measuring units (IMU), a precise and an automotive grade device.

2. DATA

Data basis for the simulation is a 21.7 km long trajectory which runs through densely built-up regions as well as along highway-like roads in the area of Hannover, Germany. The data was acquired by the Streetmapper mobile mapping system (Kremer and Hunter, 2007) to obtain a dense laser scan. From this point cloud we extracted 2658 pole-like objects, for example sign posts, street lights or tree trunks, fully automatically (Brenner, 2009). These extracted objects build the two-dimensional landmark map which is used in the simulation. The accuracy of position for every object is in the order of 12 cm (Brenner and Hofmann, 2010).

Figure 1 shows the trajectory (red line) together with the extracted objects (green dots). The objects are not distributed equally. Along highway-like roads (left side) there are only very few poles, whereas at inner-city junctions there are normally many usable objects, for example sign posts and traffic lights.

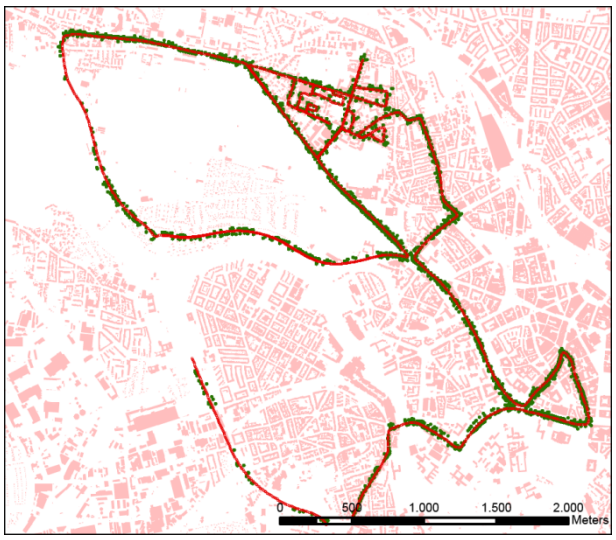


Figure 1. Trajectory (red) and extracted pole-like objects (green) which are used as reference map.

For the simulation of the positioning accuracy we used 2141 positions at a regularly spaced interval of 10 m with known coordinates and heading along the Streetmapper trajectory.

3. SIMULATION

For each of the 2141 positions along the trajectory the possible accuracy was calculated for varying opening angles (from 50° to 100° in steps of 10°) and base lengths (from 0.25 m to 2.25m in steps of 25 cm) of a simulated camera system.

The simulation is separated in three parts. The first part contains the retrieval of all visible objects within the database for each position mainly depending on the distance of the objects to the sensor to preserve a certain minimum width and the opening angle of the cameras.
The second part is the analysis of accuracies for the vehicle's position and heading, which is based on a least squares adjustment.
The third part is the analysis of accuracies using the results from the first and second part of the simulation and an additional Kalman filter.

3.1 Retrieval of Visible Objects

In the first part of the simulation all visible poles have to be determined for each position. An object is visible if its size in the image is not below a minimum size n_{px} and it lies within the overlapping field of view given by the opening angles of both cameras.
The size of the objects in the image depends on the distance between the sensor and the measured object, the focal length c, the size of the pixels on the sensor d_{px} and size of the object itself d_{obj}. A maximum distance s_{max} can be calculated as follows:

$$s_{max} = \frac{c \cdot d_{obj}}{n_{px} \cdot d_{px}} \tag{1}$$

with c = 0.006 m
d_{obj} = 0.3 m
d_{px} = 5.5·10^{-6} m.

With a given minimum size of 7 pixels only objects within a maximum distance of 46.75 m can be detected.

In addition, objects have to lie within the overlapping field of view of both cameras (Figure 2), which depends on the opening angle γ, heading κ and the camera positions $K(X_{0j}, Y_{0j})$, j = 1,2 number of camera, which are given as

$$\begin{bmatrix} X_{0j} \\ Y_{0j} \end{bmatrix} = \begin{bmatrix} X_0 \pm \frac{b}{2} \cdot \cos\kappa \\ Y_0 \pm \frac{b}{2} \cdot \sin\kappa \end{bmatrix} \tag{2}$$

where X_0, Y_0 = coordinates of centre of base line
b = base length.

For both projection centers the direction t_p for each object have to be calculated. Therefore, visible objects, which are then used for the further examination, have to match the following condition:

$$|\kappa - t_p| \leq \frac{\gamma}{2}. \tag{3}$$

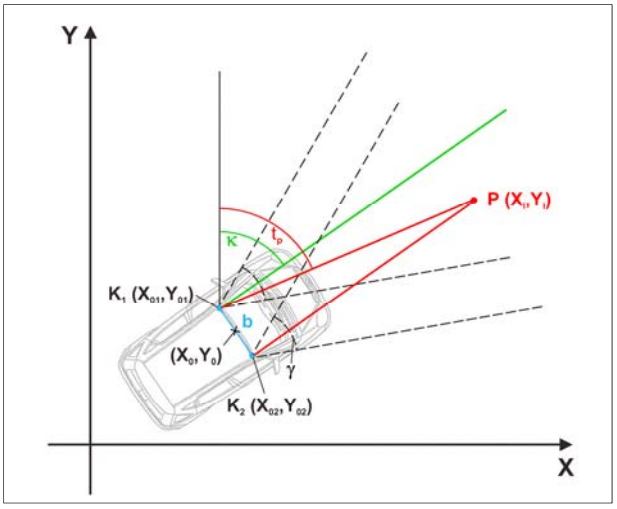


Figure 2. Visibility requirements.

3.2 Analysis of Accuracies

To determine the accuracies of the positioning (position and heading) based on stereo cameras, we used a least squares adjustment. At least two visible objects are required to calculate a position.
The functional model of the adjustment which describes the relationship between the measurements (image und corresponding object coordinates) and the unknown parameters (position and heading) are the collinearity equations. As we used a two dimensional map, the collinearity equations can be simplified and reduced to

$$x_i' = c\,\frac{r_{11}(X_i - X_0) + r_{21}(Y_i - Y_0)}{r_{12}(X_i - X_0) + r_{22}(Y_i - Y_0)} \tag{4}$$

where x_i' = image coordinate
$P(X_i, Y_i)$ = object coordinates for i-th object in the reference map
r_{ij} = elements of rotational matrix R.

In addition, we have to take into account, that the projection centers of both cameras do not lie in X_0, Y_0 but are shifted by $b/2$. Therefore, the observation equations for camera j = 1, 2 are as followed:

$$x'_{i,j} + v_{x'_{i,j}} = c\,\frac{\cos\kappa\,(X_i - X_0) + \sin\kappa\,(Y_i - Y_0) \pm \frac{b}{2}}{-\sin\kappa\,(X_i - X_0) + \cos\kappa\,(Y_i - Y_0)}. \tag{5}$$

The stochastic model contains information about the uncertainties of the observations. In the simulation object coordinates of the reference objects (standard deviation 0.1 m) and image coordinates of the measured objects (standard deviation 1/3 Pixel) are used as measurements. They are considered as fully uncorrelated, base length and focal length are considered to be accurate.

As a result, the cofactor matrix Q_{xx} of the positioning parameters contains the stochastic information of the vehicle's position:

$$Q_{xx} = (A^T \cdot P \cdot A)^{-1} \tag{6}$$

where A = design matrix
P = weight matrix.

The design matrix A provides information about the vehicle position, heading and object coordinates (number and distribution, given by reference map), thus the geometry of the setup. The weight matrix P contains the stochastic information, which is the accuracy of the image and the object coordinates.

3.3 Simulation using Filtered Positions

In a real scenario, not only individual positions would be calculated – as described before – but the positions would be filtered along the trajectories. Therefore, we simulated a second scenario, using two different types of inertial measurement units (IMU), a high precision and an automotive grade device. For the filtering, we used a standard Kalman filter based on a simplified car motion model.

The used car model is as follows

$$\begin{aligned} x_i &= x_{i-1} + dt \cdot v_{i-1} \cdot \sin \kappa_{i-1} \\ y_i &= y_{i-1} + dt \cdot v_{i-1} \cdot \cos \kappa_{i-1} \\ v_i &= v_{i-1} \\ \kappa_i &= \kappa_{i-1} \end{aligned} \tag{7}$$

where dt = time step between time i and time i-1
v_i = vehicle speed.

The transition matrix Φ is given by

$$\Phi = \begin{bmatrix} 1 & 0 & dt \cdot sin\,\kappa_{i-1} & 0 \\ 0 & 1 & dt \cdot cos\,\kappa_{i-1} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{8}$$

For the simulation we analysed the cofactor matrix Q_{xx} at time step i of the positioning parameters which is given by

$$Q^+_{xx\,(i)} = Q^-_{xx\,(i)} - K \cdot A \cdot Q^-_{xx\,(i)} \tag{9}$$

where '–' denotes the prediction and '+' denotes the update step of the filter, with

$$\begin{aligned} Q^-_{xx\,(i)} &= \Phi \cdot Q^+_{xx\,(i-1)} \cdot \Phi^{\mathrm{T}} + \mathrm{Q_{ww}} \\ K &= Q^-_{xx\,(i)} \cdot A^T \cdot (Q_{ll} + A \cdot Q^-_{xx\,(i)} \cdot A^T)^{-1} \end{aligned} \tag{11}$$

where K = Kalman gain
Q_{ll} = variance matrix of the measurements
Q_{ww} = system noise.

To filter the positions, landmark information were only used when they provided a high positioning accuracy (< 0.2 m). In all other cases only information provided by the IMU was used.

4. RESULTS

4.1 Results using Single Positions

In the first case, the simulation is based on single positions, no filtering is used along the driven path. The results show that the most important factor for the positioning accuracy is the number of objects in the field of view (Figure 3). Along the trajectory there are 2.7 to 4.6 visible objects on average depending on the opening angle (50° to 100°, respectively). With less than two visible objects, positioning is not possible. In our test area, the number of cases where we failed to retrieve the position decreases from 38.9 % to 20.6 % with increasing opening angle. Increasing the base length leads to a higher positioning accuracy, in cases where positioning is possible. For camera systems with a small opening angle, a larger base length leads to fewer visible objects based on a smaller overlap of the field of view. On average, for a typical camera system with opening angle of 100° and base length of 0.25 m, accuracies between 0.41 m (three poles visible) and 0.12 m (more than six poles visible) were achieved.

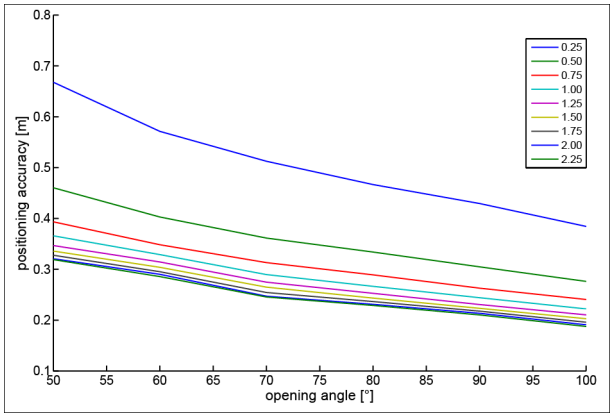


Figure 3. Positioning accuracy as a function of opening angle and base length (between 0.25 m and 2.25 m, see legend in upper right corner).

In the following examples the camera systems may have different opening angles but the same base length of 0.25 m.

In Figure 4 the positioning accuracy for two simulated camera systems are presented. Green ellipses show the positioning accuracy with 1 σ standard deviation for 100° opening angle, orange ellipses for 60° opening angle. Due to larger opening angles the overall positioning accuracy for the first case (green ellipses) is higher. For evenly distributed objects the accuracy

perpendicular to the driving direction is higher due to the model of the stereo camera. With few unfavorably distributed objects (e.g. only on one side of the road) the accuracy perpendicular to the driving direction is lower (maximum, green: 2.09 m, orange: 2.97 m). Very high accuracies (0.1 m) are achieved in both cases e.g. along an avenue lined with trees (Figure 4, upper right side).

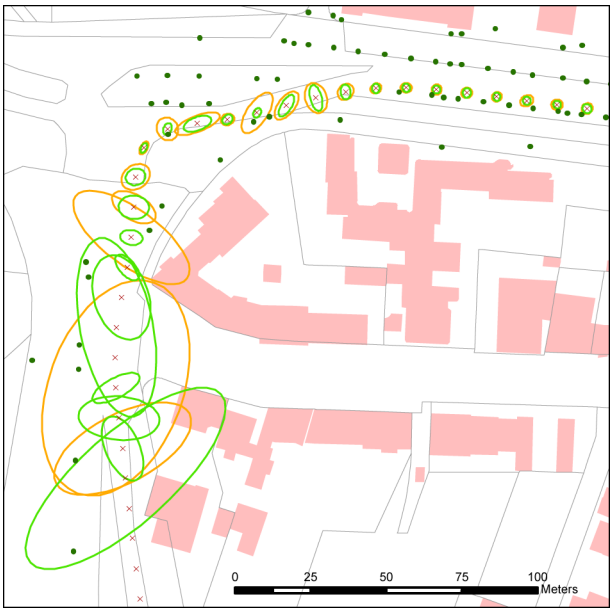


Figure 4. Positioning accuracy for single positions drawn as error ellipses (green: opening angle 100°, orange: opening angle 60°). All ellipses scaled by a factor of 20. (Red crosses: positions on trajectory, green dots: objects in reference map.)

4.2 Results using Filtered Positions

Compared to positioning based on measurements at single positions, the number of positions where positioning failed was decreased significantly when using the Kalman filter with IMU data (Figure 5 and Figure 6). With a camera system with 100° opening angle, positioning based on single positions failed in 20.6 % of the cases. Using a precise IMU (defined by σ_v = 0.02 m/s (standard deviation velocity), σ_ϕ = 0.01° (std. dev. heading), $\sigma_{\phi'}$ = 0.001°/s (std. dev. angular velocity)) the number of failures decreased to 0.2 %, where failure is defined as positioning accuracy > 20 m. Furthermore, the accuracy of calculated positions increased using a precise IMU.

Figure 5 compares the results using single positions and filtered positions based on a precise IMU. Along the highway (lower left side), very few poles lead to a high number of failed positioning, when not using a filter (green ellipses). Using a precise IMU bridges areas without a sufficient number or poorly distributed landmark objects (blue ellipses).
Using an automotive IMU (defined by σ_v = 0.1 m/s, σ_ϕ = 0. 1°, $\sigma_{\phi'}$ = 0.005°/s), the number of failures also decreased. Comparing the results (Figure 6), an automotive IMU also helps to bridge areas without landmark objects. However, for some areas with very few landmarks over a long distance the positioning accuracy was in the range of 1 m.

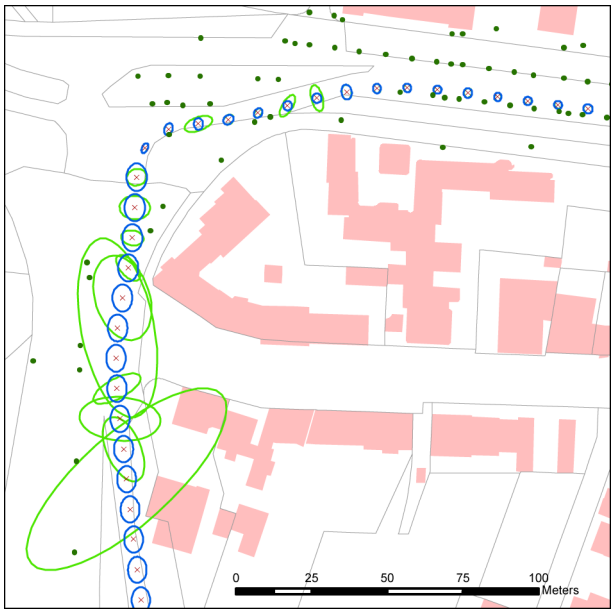


Figure 5. Positioning accuracy with filtered positions drawn as error ellipses for opening angle 100°, (blue: precise IMU, green: without IMU). All ellipses scaled by a factor of 20. (Upper right: green ellipses coincide with blue ellipses.)

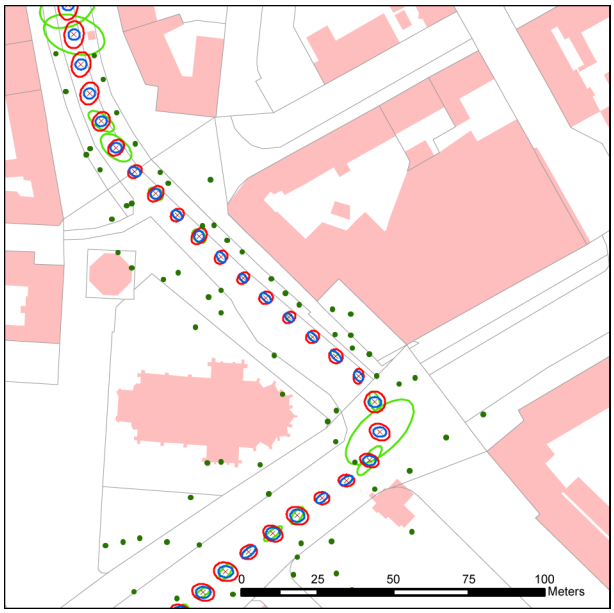


Figure 6. Positioning accuracy drawn as error ellipses for opening angle 100°, (green: without IMU, blue: precise IMU, red: automotive IMU). All ellipses scaled by a factor of 20. (Middle: green ellipses coincide with blue ellipses.)

Another important question is the maximum allowed distance between objects for positioning. Or to put it another way, how far can we drive without using any landmark objects for localization but still reach an acceptable positioning accuracy? Results are shown in Table 1, for a camera system with opening angle 100° and 0.1 m positioning accuracy at the starting point.

max. pos. error	Precise IMU	Automotive IMU
0.2 m	650 m	30 m
0.3 m	990 m	80 m

Table 1. Maximum distance without landmark update with a maximum positioning error of 0.2 m and 0.3 m.

4.3 Positioning Accuracy Maps

Based on the results of the simulation new maps of position accuracies were created (Figure 7, Figure 8, and Figure 9). These maps indicate the possible accuracy on each point along the roads. As shown here, the accuracy was only calculated along the driven path. From these maps, e.g. areas can be selected, where the positioning accuracy is lower (orange and red areas) and higher (green and yellow areas) than when using GPS. It is obvious, that in urban environments, especially at intersections, a lot of pole-like objects are present, which leads to a reliable and accurate positioning; on highways there are less of these objects, so often a positioning is impossible (Figure 7, blue areas).

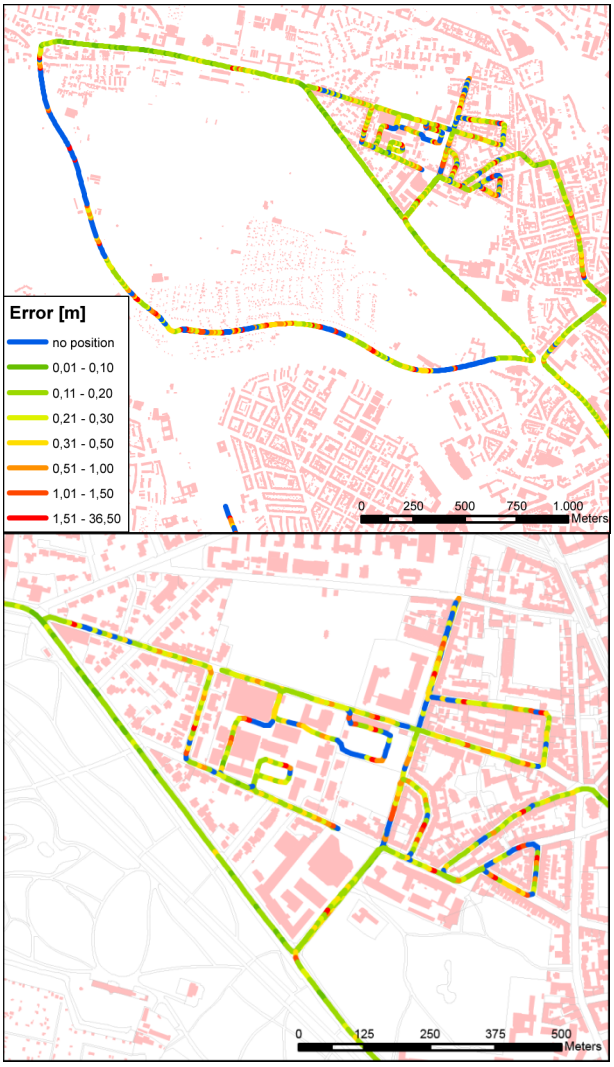


Figure 7. Map of accuracies without filter for a camera system with 100° opening angle.

Using the Kalman filter led to far better results. In Figure 8 the positioning accuracy in general is better than 0.3 m. With an automotive IMU the results were also improved (Figure 9). Even better results can be achieved when using all visible landmarks in the filter process. This is especially the case in densely-built up areas (compare Figure 7 and Figure 9, lower picture).

In general, however, GPS and landmark based navigation are somewhat complementary, as in open areas, where GPS works well, there are less possibly obtruding landmarks and vice versa.

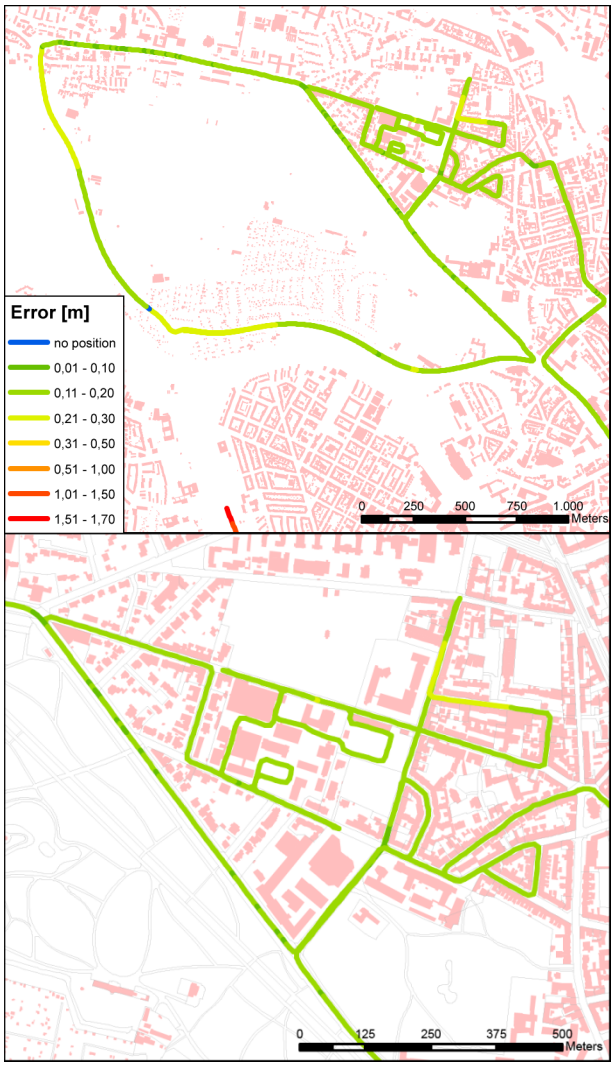


Figure 8. Map of accuracies using a precise IMU for filtering for a camera system with 100° opening angle. The accuracy along the highway (upper picture, left) depends on the direction of travel (from south to north).

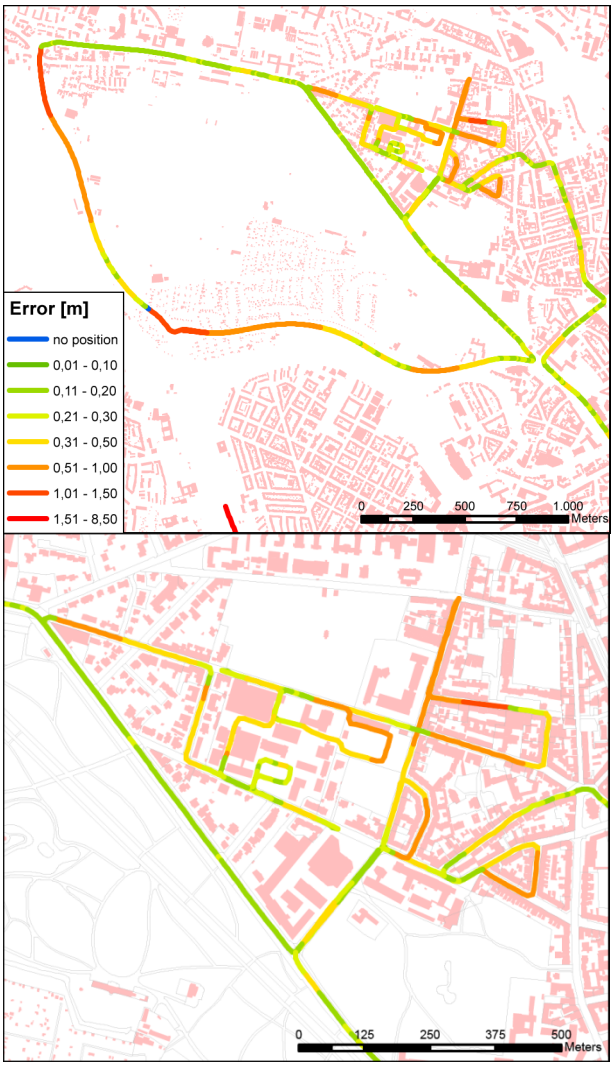


Figure 9. Map of accuracies using an automotive IMU for filtering for a camera system with 100° opening angle. The accuracy along the highway (upper picture, left) depends on the direction of travel (from south to north).

5. OUTLOOK

The results of the presented method show a high potential for localization based on landmarks. Especially at inner-city intersections, where lateral and longitudinal positioning accuracy is very important, the simulation gives promising results.

As already mentioned, positioning was not successful in all cases. In some areas especially when not using a filter the accuracy was low due to too few or unfavorably distributed objects. As shown in the second part of the simulation, using a standard Kalman filter helps to improve not only the positioning accuracy but also bridges the gap between landmarks where no positioning would be possible. Furthermore, using additional features, such as planes, which we can find in urban areas, will help to improve reliability and accuracy. We can expect, that the combination of GPS, an automotive IMU and landmarks leads to a reliable and accurate localization.

The current simulation reconstructed the positions based on stereo-reconstruction. Another simulation will be undertaken using only mono-images. Furthermore, we plan to verify the simulation with a data acquisition in a mobile mapping system.

6. REFERENCES

Brenner, C., 2009: Extraction of Features from Mobile Laser Scanning Data for Future Driver Assistance Systems, *Advances in GIScience: Proceedings of 12th AGILE Conference on GIScience, Lecture Notes in Geoinformation and Cartography*, Springer, Berlin, pp. 25-42.

Brenner, C., Hofmann, S., 2010: Evaluation of Automatically Extracted Landmarks for Future Driver Assistance Systems, *Proceedings of the Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science*, Hongkong, Vol. 38, No. 2, pp. 361-366.

Brenner, C., 2010: Vehicle localization using landmarks obtained by a LIDAR mobile mapping system, *Proceedings of the ISPRS Commission III Symposium on Photogrammetric Computer Vision and Image Analysis*, Paris, pp. 139-144.

Burgard, W., Hebert, M., 2008: *Springer Handbook of Robotics.* Springer, chapter World Modeling, pp. 853-869.

Kremer, J., Hunter, G., 2007: Performance of the streetmapper mobile lidar mapping system in real world projects. *Photogrammetric Week 2007*, Wichmann, pp. 215-225.

Weiss, T., Kaempchen, N., Dietmayer, K., 2005: Precise ego localization in urban areas using laserscanner and high accuracy feature maps. *Proc. 2005 IEEE Intelligent Vehicles Symposium*, Las Vegas, USA, pp. 284-289.

CROSS-COVARIANCE ESTIMATION FOR EKF-BASED INERTIAL AIDED MONOCULAR SLAM

Markus Kleinert[a] **and Uwe Stilla**[b]

[a]Fraunhofer IOSB, Department Scene Analysis, Gutleuthausstr. 1, 76275 Ettlingen, markus.kleinert@iosb.fraunhofer.de
[b]Technische Universität München, Photogrammetry and Remote Sensing, Arcisstrasse 21, 80333 München, stilla@bv.tum.de

KEY WORDS: Monocular SLAM, Inertial measurement system, Extended Kalman filter, Correlations, Estimation

ABSTRACT:

Repeated observation of several characteristically textured surface elements allows the reconstruction of the camera trajectory and a sparse point cloud which is often referred to as "map". The extended Kalman filter (EKF) is a popular method to address this problem, especially if real-time constraints have to be met. Inertial measurements as well as a parameterization of the state vector that conforms better to the linearity assumptions made by the EKF may be employed to reduce the impact of linearization errors. Therefore, we adopt an inertial-aided monocular SLAM approach where landmarks are parameterized in inverse depth w.r.t. the coordinate system in which they were observed for the first time. In this work we present a method to estimate the cross-covariances between landmarks which are introduced in the EKF state vector for the first time and the old filter state that can be applied in the special case at hand where each landmark is parameterized w.r.t. an individual coordinate system.

1 INTRODUCTION

1.1 Motivation

Navigation in unknown environments is often hindered by the absence of external positioning information. In the context of pedestrian navigation for instance, the GPS signal may be temporarily lost or severely disturbed due to multipath effects in urban canyons. The need to cope with such situations has motivated the research in systems which are capable of building and maintaining a map of the environment while at the same time localizing themselves w.r.t. that map. This problem is commonly referred to as simultaneous localization and mapping (SLAM). For the particular application of pedestrian navigation, a promising approach is to combine a low-cost inertial measurement unit (IMU) and a camera to an inertial aided monocular SLAM system and perform sensor data fusion in an extended Kalman filter (EKF), *cf.* (Veth and Raquet, 2006). Here, characteristically textured surface elements serve as landmarks which can be observed by the camera to build up a map while the IMU's acceleration and angular rate measurements are integrated in order to obtain a short-time accurate prediction of the camera's pose and thereby help to reduce linearization error.

1.2 Related Work

An important aspect of such monocular SLAM systems is the representation of landmarks in the filter state vector. Montiel *et al.* have proposed an inverse depth parameterization of landmarks that conforms well to the linearity assumptions made by the EKF and offers the possibility of instantly introducing new landmarks in the filter state with only one observation (Montiel et al., 2006). In the original inverse depth parameterization six additional parameters are included in the filter state for each freshly introduced landmark. To alleviate the computational burden imposed by this over-parameterization, Civera *et al.* introduce a method to transform landmarks to Cartesian coordinates once their associated covariance has sufficiently collapsed (Civera et al., 2007). Alternatively, Pietzsch proposes to initialize bundles of landmarks and to estimate only the inverse distance to the origin of the coordinate system of the camera that observed the landmarks for the first time for each landmark individually and the position and orientation of the camera coordinate system for the whole bundle (Pietzsch, 2008).

The importance of the cross-covariance terms in SLAM is stressed in a seminal work by Dissanayake *et al.* (Dissanayake et al., 2001). Julier and Uhlmann present a detailed investigation of the consistency of EKF-SLAM implementations (Julier and Uhlmann, 2001). Therein it is shown that errors in the estimated cross-covariance terms due to linearization errors lead to inconsistent estimates. A comparison of several landmark parameterizations for monocular SLAM regarding their effects on filter consistency is provided by Solà (Solà, 2010). This work also gives a detailed description of landmark initialization in monocular SLAM.

1.3 Contribution

Our approach is to parameterize each landmark in inverse depth polar coordinates w.r.t. the coordinate system of the camera at the time of its first observation. Therefore, the camera's orientation and position as well as the parameters that describe the landmark's position in the camera coordinate frame have to be stored for each landmark. However, we regard the camera's position and orientation as fix model parameters and thus only include the three parameters which describe the landmark's uncertain position in the filter state, thereby avoiding overparameterizing the landmark's position. Since the camera's position and orientation are regarded as fix model parameters, the corresponding uncertainty estimate has to be conveyed to the landmark's uncertainty estimate. In addition, the cross-covariances between the new landmark and the landmarks already present in the filter state have to be computed. This is aggravated by the fact, that the landmark coordinates in the filter state are given with respect to distinct coordinate frames, which precludes the adaption of standard error propagation methods in this case. The main contribution of our work is a method to convey the uncertainty estimate from the camera to the landmark parameters and to determine the cross-covariances between the new landmark and the parameters already present in the filter state.

2 EKF-SLAM FORMULATION

This section describes the EKF-SLAM approach employed in this work with an emphasis on the parameterization of landmarks.

2.1 Coordinate systems

The following coordinate systems are of particular interest for the derivation of the cross-covariances in sec. 2.5.2. An overview is also given in fig. 1.

The body or IMU-coordinate system $\{b\}$ that is aligned to the IMU's axes and therefore describes position and orientation of the whole sensor system. Its position and orientation are included in the filter state.

The camera coordinate system $\{c\}$. The camera's position and orientation are not part of the filter state. They can be calculated from the IMU's position and orientation by means of the camera-IMU transformation that only depends on the mechanical setup and is assumed to be fix and known in this work.

The navigation frame $\{n\}$. This is the reference frame whose x- and y- axis point north- and eastwards while its z-axis points in the direction of local gravity. We assume that the distance to the local reference frame is small compared to the curvature of the earth and therefore the direction of gravity can be considered constant during operation of the system. In this case the position of the navigation frame can be chosen arbitrarily.

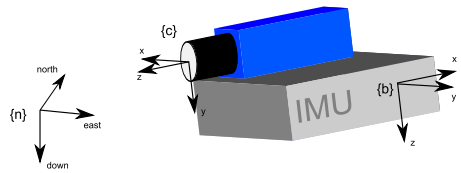


Figure 1: Overview of the coordinate systems used in this work

2.2 State parameterization

The goal is to determine the position and orientation of the body frame w.r.t. the navigation frame and a sparse map of point landmarks. Hence, the EKF state vector comprises parameters which describe the IMU's motion and biases as well as the coordinates of observed landmarks:

$$\mathbf{s}_t = \left[\underbrace{{}^n\mathbf{p}_b^{\mathrm{T}}\ {}^n\mathbf{v}_b^{\mathrm{T}}\ \mathbf{b}_a^{\mathrm{T}}\ \mathbf{q}_b^{n\mathrm{T}}\ \mathbf{b}_g^{\mathrm{T}}}_{\mathbf{s}'} \underbrace{\mathbf{Y}_1^{\mathrm{T}} \ \dots \ \mathbf{Y}_N^{\mathrm{T}}}_{\mathbf{m}} \right]^{\mathrm{T}} \quad (1)$$

Where ${}^n\mathbf{p}_b$ and ${}^n\mathbf{v}_b$ are the IMU's position and velocity w.r.t. the navigation frame, the unit quaternion $\mathbf{q}_b^n$ represents the orientation and $\mathbf{b}_a$, $\mathbf{b}_g$ are the sensor biases, which systematically disturb acceleration and angular rate measurements. For convenience, the landmark coordinates $\mathbf{Y}_i$ are subsumed in the map vector $\mathbf{m}$. Similarly, the part of the state vector that describes the motion of the IMU is denoted by $\mathbf{s}'$. In the following, estimated values are denoted with by a hat $(\hat{\cdot})$ and a tilde $(\tilde{\cdot})$ is used to indicate the error, *i.e.* the deviation between a true value $(\cdot)$ and its estimate: $(\tilde{\cdot}) = (\cdot) - (\hat{\cdot})$.

2.2.1 Error state formulation.

Since the EKF relies on a truncation of the Taylor series expansion of the measurement equation as well as the time update step after the first derivative, it can be regarded as an estimator for the state error $\tilde{\mathbf{s}}$. This is the basis for the error state formulation of the EKF which is commonly used for GPS-INS integration, *cf.* (Farrell and Barth, 1999, pp. 199-222). Therefore, the covariance matrix associated with the filter state describes the distribution of $\tilde{\mathbf{s}}$ under the assumption that the errors follow a normal distribution. It is given by

$$P = \begin{bmatrix} P_{\tilde{\mathbf{s}}',\tilde{\mathbf{s}}'} & P_{\tilde{\mathbf{s}}',\tilde{\mathbf{m}}} \\ P_{\tilde{\mathbf{m}},\tilde{\mathbf{s}}'} & P_{\tilde{\mathbf{m}},\tilde{\mathbf{m}}} \end{bmatrix}. \quad (2)$$

The error of the estimated orientation can be written in terms of the incremental orientation that aligns the estimated coordinate system with the unknown true coordinate system:

$$\mathbf{q}_b^n = \mathbf{q}(\Psi) * \hat{\mathbf{q}}_b^n \quad , \quad \mathbf{q}(\Psi) \approx \left[1 \ \tfrac{\Psi}{2}^{\mathrm{T}}\right]^{\mathrm{T}} \quad (3)$$

Where $*$ denotes quaternion multiplication.

2.2.2 Landmark parameterization.

The coordinate vector of the i-th landmark in the filter state $\mathbf{Y}_i$ describes the position of the landmark in inverse depth polar coordinates w.r.t. the coordinate frame $\{c_k\}$ of the camera at the time when the landmark was observed for the first time as illustrated in fig. 2.

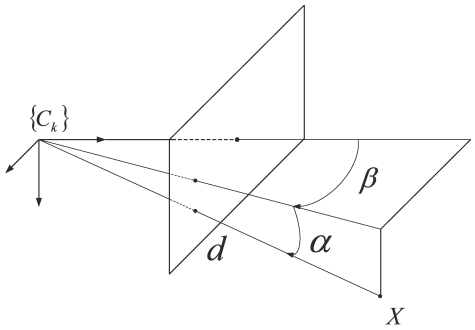


Figure 2: Landmark parameterization in inverse depth polar coordinates. X is the Cartesian coordinate vector associated with the inverse depth parameterization $Y = [\alpha\ \beta\ \rho\ ]$, with elevation angle α, azimuth β and inverse depth $\rho = 1/d$, all w.r.t. the anchor system $\{c_k\}$.

Using the camera coordinate frame $\{c_k\}$ as an anchor for the landmark therefore avoids over-parameterization in the filter state and should thus increase computational efficiency and stability during Kalman filter updates. In order to determine the position of a landmark in the reference coordinate system, the transformation from the anchor coordinate system to the reference frame is needed:

$${}^n\mathbf{X} = C_{c_k}^n \cdot \frac{1}{\rho} \underbrace{\begin{bmatrix} \cos(\alpha)\sin(\beta) \\ \sin(\alpha) \\ \cos(\alpha)\cos(\beta) \end{bmatrix}}_{\overline{\mathbf{Y}}(\mathbf{Y})} + {}^n\mathbf{p}_{c_k} \quad (4)$$

In the above equation, the direction cosine matrix $C_{c_k}^n$ describes the orientation of of the anchor system and ${}^n\mathbf{p}_{c_k}$ is its position. $\overline{\mathbf{Y}}$ is a unit vector that points in the direction of the projection ray.

For every landmark $C^n_{c_k}$ and ${}^n\mathbf{p}_{c_k}$ are therefore stored in a separate data structure because they are not part of the filter state although they vary for different landmarks.

2.3 Time update

2.3.1 Integration of inertial measurements. During the time update, the estimated state is propagated by integrating the inertial measurements. To integrate angular rate measurements ω, a quaternion that describes the incremental rotation in the body frame is formed from the angular rate measurements and subsequently used to update the orientation estimate:

$$\begin{aligned} \hat{\omega} &= \omega - \hat{\mathbf{b}}_g \\ \hat{\mathbf{q}}^n_{b,t+\tau} &= \hat{\mathbf{q}}^n_{b,t} * \mathbf{q}(\hat{\omega}) \end{aligned} \tag{5}$$

Where τ is the time interval between two consecutive inertial measurements. Acceleration measurements have to be transformed to the reference coordinate system before the gravitational acceleration can be subtracted. Then, the resulting estimate of acceleration is used to integrate velocity and position:

$$\begin{aligned} {}^n\hat{\mathbf{a}}_{b,t} &= C(\hat{\mathbf{q}}^n_{b,t}) \cdot ({}^b\tilde{\mathbf{a}}_{b,t} - \hat{\mathbf{b}}_{a,t}) + {}^n\mathbf{g} \\ {}^n\hat{\mathbf{p}}_{b,t+\tau} &= {}^n\hat{\mathbf{p}}_{b,t} + {}^n\hat{\mathbf{v}}_{b,t} \cdot \tau + \tfrac{1}{2}{}^n\hat{\mathbf{a}}_{b,t} \cdot \tau^2 \\ {}^n\hat{\mathbf{v}}_{b,t+\tau} &= {}^n\hat{\mathbf{v}}_{b,t} + {}^n\hat{\mathbf{a}}_{b,t} \cdot \tau \end{aligned} \tag{6}$$

2.3.2 Covariance propagation. The physical model described by equations 5-6 corresponds to the first order differential equation that describes the propagation of the estimation error $\tilde{\mathbf{s}}'$ for the time dependent part of the state vector:

$$\dot{\tilde{\mathbf{s}}}' = F' \cdot \tilde{\mathbf{s}}' + G' \cdot \mathbf{n} \tag{7}$$

Here, F' is determined by the physical model and $\mathbf{n}$ summarizes the white noise terms. From F' the discrete time transition matrix Φ' is computed and used thereafter to calculate the propagated error state covariance matrix $P_{t+\tau}$ as stated below:

$$\Phi' = \exp(F' \cdot \tau) \approx I_{15\times 15} + F' \cdot \tau \tag{8}$$

$$P'_{t+\tau} = \Phi' \cdot P_{\tilde{\mathbf{s}}'_t,\tilde{\mathbf{s}}'_t} \cdot \Phi'^{\mathrm{T}} + \Phi' \cdot G' \cdot Q \cdot G'^{\mathrm{T}} \cdot \Phi'^{\mathrm{T}} \cdot \tau \tag{9}$$

$$P_{t+\tau} = \begin{bmatrix} P'_{t+\tau} & \Phi' \cdot P_{\tilde{\mathbf{s}}'_t,\tilde{\mathbf{m}}_t} \\ P_{\tilde{\mathbf{m}}_t,\tilde{\mathbf{s}}'_t} \cdot \Phi'^{\mathrm{T}} & P_{\tilde{\mathbf{m}}_t,\tilde{\mathbf{m}}_t} \end{bmatrix} \tag{10}$$

In the expression above, Q is the power spectral density matrix which characterizes the noise vector $\mathbf{n}$.

2.4 Measurement update

New images are continuously triggered by the IMU as it moves along its trajectory. Therein the image coordinates of point features are extracted and tracked. The coordinates of all features extracted in one image are stacked together to form the measurement vector that is subsequently used to update the state vector.

2.4.1 Landmark observation model. The observation model describes the relationship between the observed image coordinates and the state vector entries. For this purpose, the estimate of each observed landmark in inverse depth polar coordinates w.r.t. its anchor system is first transformed to Cartesian coordinates w.r.t. the navigation frame as shown in eq. 4. Subsequently, the coordinates are transformed to the current camera coordinate system and projected on the image plane:

$$\begin{aligned} \mathbf{z} &= \mathbf{h}(\mathbf{s}) + \mathbf{v} \\ &= \pi(C^c_n \cdot (\frac{1}{\rho} \cdot C^n_{c_k} \cdot \overline{\mathbf{Y}}(\alpha,\, \beta) + {}^n\mathbf{p}_{c_k} - {}^n\mathbf{p}_c)) + \mathbf{v} \\ &= \pi(\underbrace{C^c_n \cdot (C^n_{c_k} \cdot \overline{\mathbf{Y}}(\alpha,\, \beta) + \rho \cdot ({}^n\mathbf{p}_{c_k} - {}^n\mathbf{p}_b)) + \rho \cdot {}^c\mathbf{p}_b}_{{}^c\mathbf{X}}) \\ &\quad + \mathbf{v} \end{aligned} \tag{11}$$

Where $\mathbf{z}$ are the measured image coordinates, $\mathbf{v}$ is the zero mean measurement noise, and $\pi(\cdot)$ is the projection function. Eq. 11 describes the projection of one landmark. The Jacobian of $\mathbf{h}(\mathbf{s})$ w.r.t. the entries of the state vector for landmark no. i is:

$$H_i = J_\pi \begin{bmatrix} J_{\mathbf{p}}\ 0_{3\times 6}\ J_\Psi\ 0_{3\times 3}\ 0_{3\times 3\cdot(i-1)}\ J_{\mathbf{Y}}\ 0_{3\times 3\cdot(N-i)} \end{bmatrix} . \tag{12}$$

With the derivatives J_Ψ, $J_{\mathbf{p}}$, and $J_{\mathbf{Y}}$ of ${}^c\mathbf{X}$ w.r.t. the orientation $\mathbf{q}^n_b$, the position ${}^n\mathbf{p}_b$, and the landmark $\mathbf{Y}$. Similarly to the measurement vector, the Jacobian H for the whole set of measurements is obtained by stacking the measurement Jacobians for individual landmarks.

Given the measurement model derived in this section and the prediction from sec. 2.3, an EKF update step can be performed as described in (Bar-Shalom et al., 2001, pp. 200-217).

2.5 Introduction of new landmarks

Landmarks are deleted from the filter state if they could not be observed in a predefined number of consecutive images. Whenever the number of landmarks in the state drops below a predetermined threshold, new landmarks have to be introduced. Since the standard formulation of the Kalman filter does not allow for a variable state size, the new filter state entries have to be estimated based on previous observations. The inverse depth polar coordinates for each new feature can be calculated based on the image coordinates of its last observation and the camera calibration by means of trigonometric functions:

$$\mathbf{Y}_{new} = f(p_x, p_y, k, \rho_{init}) \tag{13}$$

Where p_x, p_y are the measured image coordinates, k contains the intrinsic camera calibration parameters, ρ_{init} is an arbitrarily chosen inverse depth measurement and $f(\cdot)$ is the back projection function, which projects to a point on the projection ray through the observed image coordinates. In the following, J_f denotes the Jacobian of f w.r.t. p_x, p_y ,and ρ_{init}.

A new landmark is introduced in the Kalman filter state by augmenting the state vector with the initial estimate of the landmark's position $\mathbf{Y}_{new}$:

$$\mathbf{s}_{new} = \begin{bmatrix} \mathbf{s}'^{\mathrm{T}}\ \mathbf{m}^{\mathrm{T}}\ \mathbf{Y}^{\mathrm{T}}_{new} \end{bmatrix}^{\mathrm{T}} \tag{14}$$

In addition the covariance matrix has to be augmented with the new cross-covariance terms and the covariance of the new landmark:

$$P_{new} = \begin{bmatrix} P_{\tilde{\mathbf{s}}',\tilde{\mathbf{s}}'} & P_{\tilde{\mathbf{s}}',\tilde{\mathbf{m}}} & P_{\tilde{\mathbf{s}}',\tilde{\mathbf{Y}}_{new}} \\ P_{\tilde{\mathbf{m}},\tilde{\mathbf{s}}'} & P_{\tilde{\mathbf{m}},\tilde{\mathbf{m}}} & P_{\tilde{\mathbf{m}},\tilde{\mathbf{Y}}_{new}} \\ P_{\tilde{\mathbf{Y}}_{new},\tilde{\mathbf{s}}'} & P_{\tilde{\mathbf{Y}}_{new},\tilde{\mathbf{m}}} & P_{\tilde{\mathbf{Y}}_{new},\tilde{\mathbf{Y}}_{new}} \end{bmatrix} \quad (15)$$

2.5.1 Conventional approach. A commonly used method to calculate the initial landmark estimate and the associated covariance entries is to define an inverse observation function

$$\mathbf{Y}_{new} = \mathbf{g}(\mathbf{s}', \mathbf{z}, \rho) \quad (16)$$

that depends on one or more measurements $\mathbf{z}$, the sensor system's position and orientation in $\mathbf{s}'$ as well as on predefined parameters like ρ. Let $J_{s'}$, J_z, and J_ρ be the derivatives of $\mathbf{g}(\cdot)$ w.r.t. $\mathbf{s}'$, $\mathbf{z}$, and ρ. The sought-after covariance entries can then be approximated as follows (Solà, 2010):

$$P_{\tilde{\mathbf{Y}}_{new},\tilde{\mathbf{s}}'} = J_{s'} P_{\tilde{\mathbf{s}}',\tilde{\mathbf{s}}'} \quad (17)$$

$$P_{\tilde{\mathbf{Y}}_{new},\tilde{\mathbf{m}}} = J_{s'} P_{\tilde{\mathbf{s}}',\tilde{\mathbf{m}}'} \quad (18)$$

$$P_{\tilde{\mathbf{Y}}_{new},\tilde{\mathbf{Y}}_{new}} = J_{s'} P_{\tilde{\mathbf{s}}',\tilde{\mathbf{s}}'} J_{s'}^{\mathrm{T}} + J_z R J_z^{\mathrm{T}} + J_\rho \sigma_\rho^2 J_\rho^{\mathrm{T}} \quad (19)$$

Where σ_ρ^2 is the variance of the initial inverse depth estimate and R is the measurement noise covariance matrix.

2.5.2 Proposed method. The scheme presented in the previous section is not directly applicable to landmark parameterizations as described in sec. 2.2.2. In this case function $f(\cdot)$ from eq. 13 takes the role of $g(\cdot)$ in the above equations. The problem is, that $f(\cdot)$ does not depend on the system's position and orientation. Thus, the Jacobian $J_{s'}$ and with it the cross-covariances $P_{\tilde{\mathbf{Y}}_{new},\tilde{\mathbf{s}}'}$, $P_{\tilde{\mathbf{Y}}_{new},\tilde{\mathbf{m}}}$ become zero. As a result, the uncertainty of the position and orientation estimate of the sensor system will be neglected when eq. 17- 19 are applied.

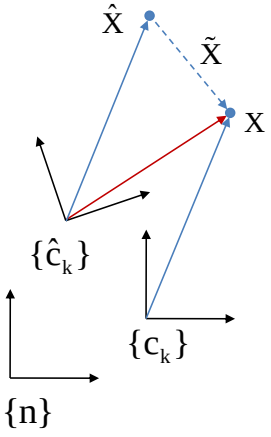


Figure 3: Position of a feature expressed in the true camera coordinate system c_k and the estimated camera coordinate system $\hat{c}_k$.

Fig. 3 depicts the situation when a landmark's position estimate in Cartesian coordinates w.r.t. the reference frame $\{n\}$ is computed based on an erroneous estimate $\{\hat{c}_k\}$ of the camera's position and orientation. The blue arrows indicate the landmark position estimates in the true camera coordinate system $\{c_k\}$ and in the estimated camera coordinate system $\{\hat{c}_k\}$ that are consistent with the observed feature locations. Note, that the vectors $\hat{\mathbf{X}}$, $\mathbf{X}$ expressed in camera coordinates are identical in case of a perfect measurement. The red arrow marks the correct landmark estimate in the erroneously estimated camera coordinate system $\{c_k\}$. The key idea behind our approach is to express the landmark position error $\tilde{\mathbf{X}} = \mathbf{X} - \hat{\mathbf{X}}$ (the dashed arrow) in the estimated camera coordinate frame in terms of the unknown transformation between the true and the estimated camera coordinate systems and to use this transformation to propagate the error from the camera coordinate system estimate to the landmark position estimate and to calculate the cross-covariance entries.

In the ensuing derivation, the orientation error model described in sec. 2.2.1 will be used. The transformation between the true coordinate system $\{c_k\}$ and the estimated coordinate system $\{\hat{c}_k\}$ can be written in terms of a rotation matrix $C_{c_k}^{\hat{c}_k}$ and the position ${}^{\hat{c}_k}\mathbf{p}_{c_k}$. The rotation matrix $C_{c_k}^{\hat{c}_k}$ depends on the Rodrigues vector that is defined in eq. 3:

$$C_{c_k}^{\hat{c}_k} = C(\Psi_{c_k}^{\hat{c}_k}) \approx (I - \left\lfloor \Psi_{c_k}^{\hat{c}_k} \right\rfloor_\times) \quad (20)$$

$$\Psi_{c_k}^{\hat{c}_k} = \hat{C}_n^{c_k} \cdot \Psi \quad (21)$$

Where $\hat{C}_n^{c_k} = C_n^{\hat{c}_k} = C_{\hat{n}}^{c_k}$ is the rotation matrix calculated from the estimated orientation of the sensor system and the IMU-camera calibration and $\Psi_{c_k}^{\hat{c}_k}$ is the orientation error expressed in the estimated camera coordinate system. With this an expression for the landmark error in Cartesian coordinates can be derived where the index k, which is used to mark the anchor coordinate system for a landmark, is omitted for brevity:

$$\begin{aligned} {}^{\hat{c}}\tilde{\mathbf{X}} &= {}^{\hat{c}}\mathbf{X} - {}^{\hat{c}}\hat{\mathbf{X}} \\ &= C_c^{\hat{c}} \cdot {}^{c}\mathbf{X} + {}^{\hat{c}}\mathbf{p}_c - {}^{\hat{c}}\hat{\mathbf{X}} \\ &= C_c^{\hat{c}} \cdot {}^{c}\mathbf{X} + C_n^{\hat{c}} \cdot ({}^{n}\mathbf{p}_b - {}^{n}\mathbf{p}_{\hat{b}}) + {}^{c}\mathbf{p}_b - C_c^{\hat{c}} \cdot {}^{c}\mathbf{p}_b - {}^{\hat{c}}\hat{\mathbf{X}} \\ &= C_c^{\hat{c}} \cdot ({}^{c}\mathbf{X} - {}^{c}\mathbf{p}_b) - {}^{\hat{c}}\hat{\mathbf{X}} + {}^{c}\mathbf{p}_b + C_n^{\hat{c}} \cdot ({}^{n}\mathbf{p}_b - {}^{n}\mathbf{p}_{\hat{b}}) \\ &\approx C_c^{\hat{c}} \cdot ({}^{\hat{c}}\hat{\mathbf{X}} - {}^{c}\mathbf{p}_b) - {}^{\hat{c}}\hat{\mathbf{X}} + {}^{c}\mathbf{p}_b + C_n^{\hat{c}} \cdot ({}^{n}\mathbf{p}_b - {}^{n}\mathbf{p}_{\hat{b}}) \end{aligned} \quad (22)$$

In eq. 22 the approximation ${}^{c}\mathbf{X} \approx {}^{\hat{c}}\hat{\mathbf{X}}$ is used, thereby assuming that the main error is caused by the erroneous estimate of the coordinate system, *cf.* fig. 3. Using the small angle approximation from eq. 20, ${}^{\hat{c}}\tilde{\mathbf{X}}$ can be written as a linear function of the errors of the estimated state:

$$\begin{aligned} {}^{\hat{c}}\tilde{\mathbf{X}} &= (I + \left\lfloor \Psi_{c_k}^{\hat{c}_k} \right\rfloor_\times) \cdot ({}^{\hat{c}}\hat{\mathbf{X}} - {}^{c}\mathbf{p}_b) - {}^{\hat{c}}\hat{\mathbf{X}} + {}^{c}\mathbf{p}_b + C_n^{\hat{c}} \cdot ({}^{n}\mathbf{p}_b - {}^{n}\mathbf{p}_{\hat{b}}) \\ &= \left\lfloor {}^{c}\mathbf{p}_b - {}^{\hat{c}}\hat{\mathbf{X}} \right\rfloor_\times \cdot C_n^{\hat{c}} \cdot \Psi + C_n^{\hat{c}} \cdot {}^{n}\tilde{\mathbf{p}}_b \end{aligned} \quad (23)$$

This is the sought relationship between the errors of the current orientation and position estimates for the sensor system and the error of the newly introduced landmark in Cartesian coordinates w.r.t. the current camera coordinate system. It only depends on entities, which are either estimated or known a-priori, like the the position of the camera in the body coordinate system. The partial derivatives of the landmark coordinates w.r.t. the IMU's position and orientation follow directly from eq. 23:

$$\frac{\partial^{c_k}\mathbf{X}}{\partial\Psi} = \left[{}^{c}\mathbf{p}_b - {}^{\hat{c}}\hat{\mathbf{X}}\right]_\times \cdot C_n^{\hat{c}} \tag{24}$$

$$\frac{\partial^{c_k}\mathbf{X}}{\partial^n\mathbf{p}_b} = \hat{C}_n^{\hat{c}} \tag{25}$$

Given the partial derivatives 24-25, a new landmark can be introduced to the filter state by following the subsequent steps:

1. Calculate the inverse depth polar coordinates $\mathbf{Y}$ and the Cartesian coordinates ${}^{c_k}\mathbf{X}$ for the new landmark given the observation and an arbitrarily chosen inverse depth estimate ρ.

2. Calculate the partial derivative $\partial\mathbf{Y}/\partial^{c_k}\mathbf{X}$, which describes how the inverse depth polar coordinates vary with ${}^{c_k}\mathbf{X}$.

3. Determine $J_{s'}$:

$$J_{s'} = \frac{\partial\mathbf{Y}}{\partial^{c_k}\mathbf{X}} \cdot \left[\frac{\partial^{c_k}\mathbf{X}}{\partial^n\mathbf{p}_b}\ 0_{3\times6}\ \frac{\partial^{c_k}\mathbf{X}}{\partial\Psi}\ 0_{3\times3}\right] \tag{26}$$

4. Use eqs. 17-19 to calculate the missing covariance matrix entries and augment the filter state according to eqs. 14 and 15.

3 RESULTS AND DISCUSSION

3.1 Experimental setup

The cross covariance estimation algorithm derived in sec. 2.5.2 was compared against a naive landmark introduction method that simply discards the cross-correlations in a number of simulation runs. For the simulation a reference trajectory was defined by two C^2 splines. One spline determines the viewing direction while the other spline describes the position of the IMU. The second derivative of this spline provides acceleration measurements whereas angular rate measurements are derived from the differential rotation between two sampling points. In addition, image measurements were generated by projecting landmark positions onto the image plane.

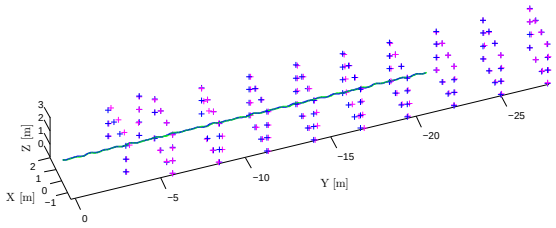
Artificial white Gaussian noise was added to all measurements. Its variance was chosen to resemble the characteristics of a good tactical grade IMU with $0.47^\circ/\sqrt{\text{h}}$ angular random walk and $0.0375\text{m}/(\text{s}\sqrt{\text{h}})$ velocity random walk parameters. The artificial noise added to the image coordinates had a standard deviation of 0.1 Pixel. Though the IMU measurement biases are modeled as random walk processes, their values stayed fix for the duration of the simulation. However, the biases were also estimated during the simulation runs, *i.e.* their initial covariance and their process noise power spectral density were initialized with realistic values. The state was initialized with the true values from the reference trajectory after a standstill period of 3.2 seconds. Deviating from the physical model described in sec. 2.3.2 a small amount of pseudo-noise was added to the diagonal elements of the covariance matrix for the landmark position estimates.

The simulation provides ground truth for the position and orientation of the IMU but not for the estimated uncertainty (the covariance matrix). Therefore, the normalized estimation error squared (NEES) is used as a measure of filter consistency for the comparison of the two methods, *cf.* (Bar-Shalom et al., 2001, p. 165):

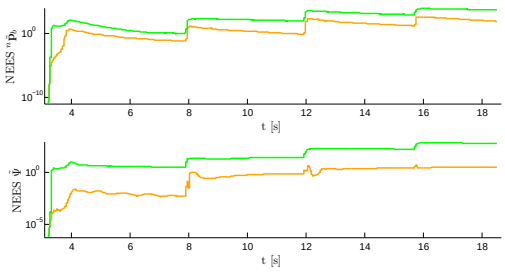
$$NEES\ {}^n\tilde{\mathbf{p}}_b = {}^n\tilde{\mathbf{p}}_b^{\mathrm{T}} P^{-1}_{{}^n\tilde{\mathbf{p}}_b,{}^n\tilde{\mathbf{p}}_b}\ {}^n\tilde{\mathbf{p}}_b \tag{27}$$

$$NEES\ \tilde{\Psi} = \tilde{\Psi}^{\mathrm{T}} P^{-1}_{\tilde{\Psi},\tilde{\Psi}}\tilde{\Psi} \tag{28}$$

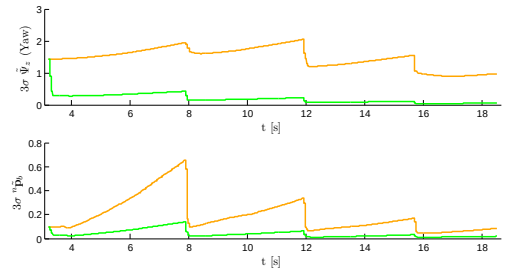
It is also interesting to investigate the covariance bounds for position and orientation errors. Since no measurements are available that relate the sensor system to the navigation frame, aside from the implicitly measured gravity vector, the uncertainty of the position and yaw angle estimates should not fall below their initial estimates.



(a) Simulated hallway trajectory



(b) NEES for position and orientation estimates



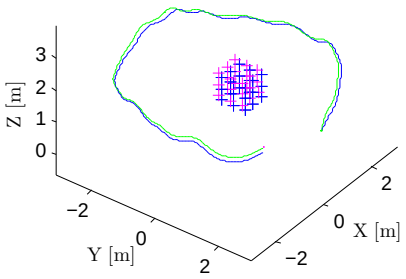
(c) Yaw angle and position covariance bounds

Figure 4: Hallway sequence. Results for one simulation run are shown in Fig. 4(a). Blue: Reference trajectory and reference landmark positions, Pink: Estimated landmark positions, Green: Estimated trajectory. Fig. 4(b) shows the NEES averaged over 25 Monte-Carlo simulation runs. Orange: With cross-covariance estimation for new landmarks. Green: without cross-covariance estimation. Notice the log scale in the NEES-plots. Fig. 4(c) compares the estimated covariance bounds for yaw angle and position estimates. Orange: With cross-covariance estimation for new landmarks. Green: without cross-covariance estimation.

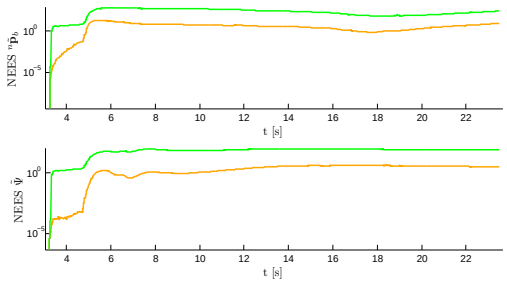
3.2 Results

Fig. 4 presents the results for a simulated trajectory that imitates a walk along a long hallway. During the simulation run, landmarks go out of sight and are replaced by newly initialized landmarks. A comparison of the NEES plots shows that estimating the cross-covariances with the proposed method indeed yields more consistent filter estimates. However, the initialization of new landmarks after 8,12, and 16 seconds goes along with a considerable drop in the uncertainty estimate and an increasing NEES. This is probably because the linearization points used to calculate the derivatives for cross-covariance estimation deviate increasingly from the ground truth during the simulation run.

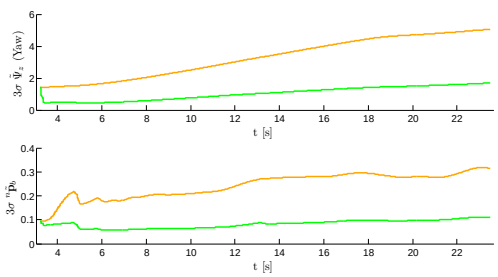
By contrast, fig. 5 shows the evaluation of a trajectory around a cube. Here, the camera's principal axis always points in the direction of the cube so that all landmarks are visible during the whole run, *i.e.* the cube is completely transparent. Thus, landmarks are initialized only once at the beginning of the run when the filter is initialized with the true parameters from the ground truth. Apparently this results in considerable more consistent estimates. In particular, the uncertainty estimate never falls below its initial value when the proposed cross-covariance estimation method is applied.



(a) Simulated trajectory around a cube



(b) NEES for position and orientation estimates



(c) Yaw angle and position covariance bounds

Figure 5: Cube sequence. See fig. 4 for details.

4 CONCLUSIONS

In this work we study the effects of inter-landmark cross-covariance estimation for an EKF-based inertial aided monocular SLAM system. In particular, we describe how the cross-covariances between new features and existing filter state entries may be computed for the special case when the landmarks are parameterized w.r.t. coordinate systems whose position and orientation is also uncertain. This situation naturally arises when parameterizing features with inverse depth polar coordinates w.r.t. the camera in which they were observed for the first time. Using simulation runs, we show that neglecting the cross-covariances for freshly inserted features results in a systematic underestimation of the filter state uncertainty and that this effect may be mitigated with the proposed algorithm.

REFERENCES

Bar-Shalom, Y., Li, X. R. and Kirubarajan, T., 2001. Estimation with Applications to Tracking and Navigation. John Wiley & Sons, Inc.

Civera, J., Davison, A. and Montiel, J., 2007. Inverse depth to depth conversion for monocular slam. In: Proceedings of ICRA.

Dissanayake, G., Newman, P., Clark, S., Durrant-Whyte, H. and Csorba, M., 2001. A solution to the simultaneous localization and map building (slam) problem. IEEE Transactions on Robotics and Automation 17, pp. 229–241.

Farrell, J. and Barth, M., 1999. The Global Positioning System & Inertial Navigation. McGraw-Hill.

Julier, S. and Uhlmann, J., 2001. A counter example to the theory of simultaneous localization and mapping. In: Proceedings of ICRA 2001.

Montiel, J., Civera, J. and Davison, A., 2006. Unified inverse depth parameterization for monocular slam. In: Robotics Science and Systems.

Pietzsch, T., 2008. Efficient feature parameterisation for visual slam using inverse depth bundles. In: Proceedings of BMVC.

Solà, J., 2010. Consistency of the monocular ekf-slam algorithm for three different landmark parameterizations. In: Proceedings of ICRA 2010.

Veth, M. and Raquet, J., 2006. Fusion of low-cost imaging and inertial sensors for navigation. In: Proceedings of the Institute of Navigation GNSS.

ACCURACY EVALUATION FOR A PRECISE INDOOR MULTI-CAMERA POSE ESTIMATION SYSTEM

C. Götz, S. Tuttas, L. Hoegner, K. Eder, U. Stilla

Photogrammetry and Remote Sensing, Technische Universitaet Muenchen, Arcisstraße 21, 80333 Munich, Germany
carsten.goetz@bv.tum.de, stilla@tum.de

Working Groups III/1, III/5

KEY WORDS: Close Range, Industrial, Pose Estimation, Accuracy

ABSTRACT:

Pose estimation is used for different applications like indoor positioning, simultaneous localization and mapping (SLAM), industrial measurement and robot calibration. For industrial applications several approaches dealing with the subject of pose estimation employ photogrammetric methods. Cameras which observe an object from a given point of view are utilized as well as cameras which are firmly mounted on the object that is to be oriented. Since it is not always possible to create an environment that the camera can observe the object, we concentrate on the latter option. A camera system shall be developed which is flexibly applicable in an indoor environment, and can cope with different occlusion situations, varying distances and density of reference marks. For this purpose in a first step a conception has been designed and a test scenario was created to evaluate different camera configurations and reference mark distributions. Both issues, the theoretical concept as well as the experimental setup are subject of this document.

1. INTRODUCTION

1.1 Motivation

Nowadays quality control in the view of geometric accurateness is an essential task in industrial manufacturing. According to a survey accomplished in 2009 by the Fraunhofer-Allianz Vision (Sackewitz, 2009), about 85% of German companies in the automotive industry employ 3D measurement techniques for different tasks. About 40% of them are using optical measurement systems or systems combined with them. For achieving highly accurate results in an absolute coordinate system, the pose of the measurement device must be known very precisely. Subject to certain conditions, determining these orientation parameters is feasible without greater efforts by means of additional sensor systems. An operable but cost-intensive solution would be the use of a laser tracker in combination with a six degrees of freedom (6DOF) tracking device. If the measurement task cannot comply with the necessary conditions, conventional solutions will not be applicable.

In literature the task of determining the orientation of an object is related to the problem of pose estimation. During the last two decades many articles have been published about photogrammetric approaches dealing with pose estimation. It is subject of various applications ranging from pose detection of persons, localization of vehicles to industrial applications. Only a brief choice will be mentioned here to show the diversity in which pose estimation could be used.

Hahn et al. (2010) present a method for tracking the spatio-temporal 3D pose of a human hand-forearm. Willert (2010) developed an approach to determine a person's position within a building using an image taken by a cell phone. An overview about optical systems for indoor self-positioning in general is given in Mautz & Tilch (2010). Their own system uses a set of projected laser points which are detected by a digital camera. Beyond indoor applications Muffert et al. (2010) investigated the quality of the spatial trajectory of a mobile survey vehicle from images recorded by an omnidirectional camera system. Another approach that exploits image sequences is introduced by Chen & Schonfeld (2010). To estimate an object's pose from multiple cameras they firstly derived a solution for only one camera employing a feature based method and extended it for multiple cameras.

Another field of research is the so called 'simultaneous localization and mapping' (SLAM). Early work pursuing different methods has been published by Facchinetti et al. (1995) and Wells et al. (1996). More recent developments can be found in Mouragnon (2006), who improved accurateness and speed of the localization and mapping of a moving vehicle by a local bundle adjustment. In addition Lemaire (2007) conducted a comparison between an algorithm that relies on monocular vision and a solution using stereovision observations. In Linkugel & Bobey (2010) a stereovision approach employing the Speeded Up Robust Feature Algorithm (SURF) is used for detection of artificial and natural landmarks. Gupta & Jarvis (2010) showed the feasibility of a localisation system for a mobile robot based on a camera with optics providing a field of view of 180°.

Further publications concentrate on industrial measurement and robot calibration. For the latter photogrammetric approaches have already been proposed in the 1990s (Albada et al., 1995; Maas, 1997). Hefele & Brenner (2001) examined the calibration done by a target board mounted to the robot and a camera placed at a fixed position as well as vice versa. Aside of applications where the industrial robot itself is of interest, pose estimation is used in conjunction with measurement devices (Sahrhage et al., 2006; Aicon3d, 2011).

Throughout all these different fields like indoor applications, SLAM and industrial applications a contradiction exists between the demand for large measurement volumes and high accuracy. So the solution for the estimated pose is a compromise with respect to the intended application.

We aim to develop a camera system for industrial applications which is able to determine its pose with submillimeter accuracy but which is useable within a range of several meters. For that purpose we plan to establish a coordinate system composed of precisely measured coded reference targets. Possible cases of application can be seen for example in processes like industrial assembly, quality control, the foresaid robot calibration or the precise alignment of tools, which are fixed on such a pose estimation system. This paper describes a concept as well as the description of a first test setup whose results will allow the evaluation of different adjustment models.

1.2 Related Work

One possibility to classify the approaches is to group them in terms of the state of motion of the pose determining sensor.
The first type would be any kind of external sensor, which operates at a fixed position observing an object's pose in a constant direction. For example Sahrhage et al. (2006) follows this concept utilizing a stereovision camera system. Schütze et al. (2010) propose a camera frame consisting of four sensors observing an active target carried by a robot. Their approach could improve the absolute positioning accuracy of the robot by a factor of 20 compared to the accuracy without any further means. The disadvantage of methods which estimate an object's pose by an external sensor is that it can hardly cope with randomly occurring occlusions, except a high redundant number of sensors is used.
The second type of pose estimating sensors can be seen in systems mounted directly on the object for which the pose has to be determined. Muffert et al. (2010) has mounted the omnidirectional camera system Ladybug 3 (Point Grey, 2011) on a mobile survey vehicle. The computation of the spatial trajectory is done from the parameters of orientation and position calculated from image sequences. As they do not bring in any pass-point information, an absolute orientation cannot be carried out.
In the field of industrial applications, the handheld probe Aicon ProCam (Aicon3d, 2011) can also be mentioned. But there is the limitation, that its measurement tip is only used to measure the coordinates of a single point and not a complete pose. The precision is specified to be 0,1mm + 0,1mm/m depending on the distance to the reference board.
Luhmann (2009) carried out an investigation on the theoretical precision of the measurement of position and orientation of an object in 3D space with respect to a reference system using a single camera. He uses two space resections, one for the transformation between camera and reference field and one for the transformation between camera and (moving) object. Although his explanations are related to the first type (sensor not in motion), one could imagine the camera being fixed to the moving object and adapt the procedure.
Frahm et al. (2004) developed an approach to estimate the pose of a multi-camera system. They assume for the system that the cameras have fixed orientations and translations between each other. Their method is applicable even in the case of non-overlapping views of the cameras. A technique to estimate the relative orientation of all the mounted cameras directly from the image sequence itself is also given.
A similar approach is that of Muhle et al. (2008). They describe a process to determine the relative orientation of two rigidly connected stereo camera systems. But both of the last-mentioned methods do not concentrate on determining an accurate pose under industrial conditions, which is the aim of our approach.

None of the approaches mentioned above does simultaneously meet the requirements for a system that can achieve a high accuracy despite randomly occurring occlusions at relatively low costs.

1.3 Overview

In section 2 of this article the demanded properties for the future pose estimation system are derived. First we sketch a coarse application scenario, and then we give an overview of the steps that have to be executed during the process. Section 3 deals with a first experimental setup, which has been designed to investigate the potential of the utilized hardware and to give verification for theoretically derived results. In section **Fehler! Verweisquelle konnte nicht gefunden werden.** the steps for a following data processing are drafted. Section 5 contains a brief discussion of the experiment. At least, in section 6, a perspective to future work and studies in this field is mentioned.

2. SYSTEM SETUP

2.1 Requirements

The photogrammetric pose estimation sensor to be developed should utilise coded reference targets to determine its absolute pose within a global coordinate system. The main principle of the approach is illustrated in Figure 1. The sensor shall be flexibly applicable in an indoor environment and shall cope with different occlusion situations, varying distances and density of reference marks. To this we examine possible configurations of such a measurement device and a setup of reference targets for evaluating the accuracy potential.

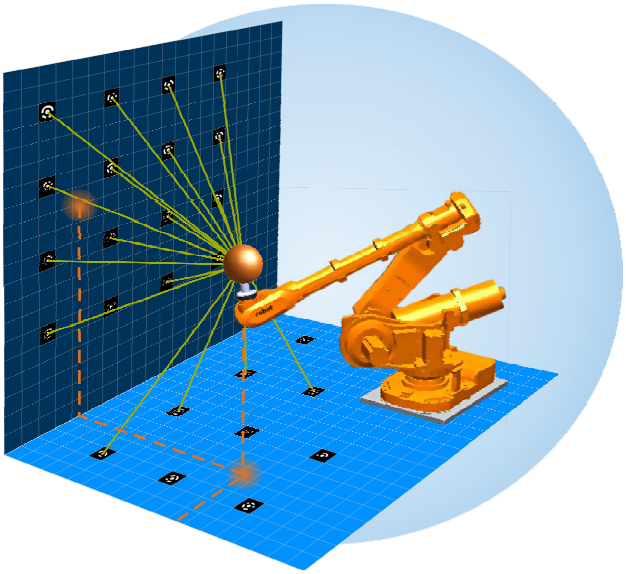


Figure 1. Pose estimation by a camera system observing surrounding reference targets.

A basic condition is, that the position at which an exact pose has to be determined, is controlled by other processes. That means that the position can be far from being optimal to determine the absolute pose in a best way. It is also not assured that there is a homogenous distribution of reference marks available since there could be limited possibilities to place them in the surrounding area. So our measurement device has to determine a precise pose, even in the case of larger occluded parts in one or more of the images.

The intended accuracy should be better than 0.2 mm for the absolute positioning within the established coordinate system of the tool used by a robot or in a more general view the robot's tool centre point. As mentioned before, this accuracy has to be achieved during suboptimal measurement conditions (occlusions, different light situations) and within a typical measurement volume of about 8x8x5 m³.
These requirements reflect a scenario of an industrial robot at a production line equipped with tools that need to be applied very precisely or carrying any other type of measurement device for quality control. In both situations the object to be worked at can have a different shape, can be located at varying places or even be in motion and occlusions of reference targets can occur very often.
As one cannot predict the visibility of reference marks (in cause of occlusion), the selection of hardware components is a compromise between the field of view and the accuracy of the target measurement within the acquired images.
Concerning the optics, a short focal length increases the chance to image any of the coded reference targets but also reduces the size of the target in an image. The latter will lead to a worse image measurement or even to the circumstance, that a target cannot be measured at all. Within this context also fisheye lenses must be evaluated in a further step, as their hemispherical view would open up a construction of a sensor which gives nearly a full field of view by comparatively low costs. Beyond that, the optics should be of compact dimensions, of low weight and offer the possibility to fix aperture and focus in a stable manner.
Concerning the requirements to the camera it can be said, that a high resolution will be helpful to detect targets and their codes respectively. But it must be considered, that the sensor size also affects the field of view. In addition, for the use in an industrial environment a compact and robust camera body is also important. Especially dust and - depending on the use of the system - affecting acceleration forces must not impair the cameras.
These considerations would lead to a compact multi-camera system which can provide a precise position, even if one (or more) cameras are not able to see reference marks. The advantage compared to approaches using an external pose estimation system is, that there is a chance to compute the pose even in situations, where the line of sight is blocked by any other object.

2.2 Concept

In a first step, the pose of the sensor is estimated only by the use of coded reference targets, whose coordinates are precisely known in a global coordinate system. At any position within that reference field, each of the n used cameras acquires an image (Figure 2, see A). The interior orientation parameters of each camera are assumed to be known and stable over the time that is needed for one specific task. Furthermore the cameras are fixed on a stable platform, so that their relative orientation can be determined in advance and will remain constant.
The circular centres of the reference targets are measured automatically. Since the orientation of all targets is known, a correction can be added to the image coordinates of a target centre if the target is too close to the sensor. This reduces errors caused by the divergence of the true target centre and the centre of the measured ellipse (Dold, 1996). To compute an adequate set of exterior orientation parameters for each image, a closed form solution for space resection will be applied (Rohrberg, 2009) in conjunction with the RANSAC algorithm (Fischler & Bolles, 1981) depending on the total number of recognized targets.
With the obtained values, which are considered as initial values, a check can be made to verify the target codes recognized. If a code cannot be validated, the affected target has to be excluded from further processing. If a specific probability exists, that a certain other code belongs to that target, it depends on the remaining detected targets, if that code is assigned to the doubtful target or if the target is just rejected.

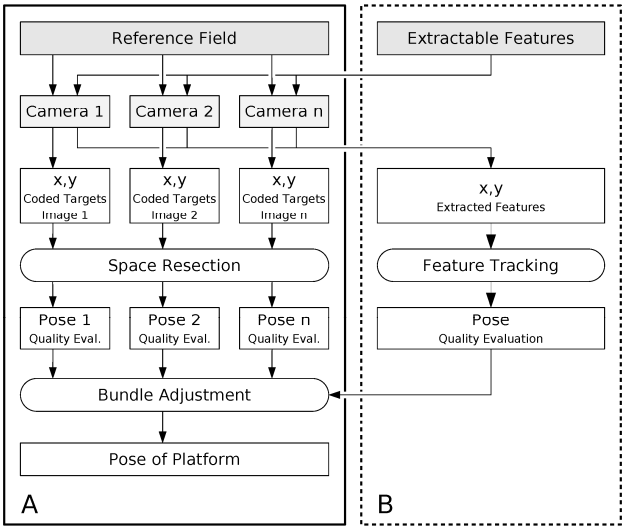


Figure 2. Process overview. System at stage 1 (see A) which will be extended to a stage 2 to use feature tracking for supporting pose estimation (B).

Afterwards the parameters are refined by an adjustment approach using only the validated subset of recognized reference marks. This yields information about the variances of the determined parameters and can be used for a quality evaluation. If a camera is not able to be oriented with a sufficient accuracy, its measurements can be excluded from the final bundle adjustment.
In a second step, the system will be extended to improve the stability of the results in difficult situations (Figure 2, B). It is intended to support the pose estimation process by tracking extractable features if not enough coded reference marks are visible. In such a situation a process needs to be introduced to support the decision whether the orientation of a camera based on feature tracking can enhance the final result or if the information must be rejected.

3. FIRST EXPERIMENTAL SETUP

3.1 Purpose of the experiment

The purpose of the experiment is to obtain a dataset from which the pose of two cameras can be computed in various combinations. This will reveal the potential for determination of the exterior orientation of the selected hardware with special attention to the translations and point out the advantages of different configurations. With a configuration the arrangement of the cameras, precisely said their distance and their relative orientation is meant. A second aspect is the verification of computations which shall simulate the same configuration as the real tests.

3.2 Experimental Setup

Two very compact industrial cameras are mounted on a platform (Figure 3), which is moveable along a linear slide rail. The change of the coordinates in linear direction can be verified by the measurement with a laser interferometer as a reflector is also mounted on that platform.

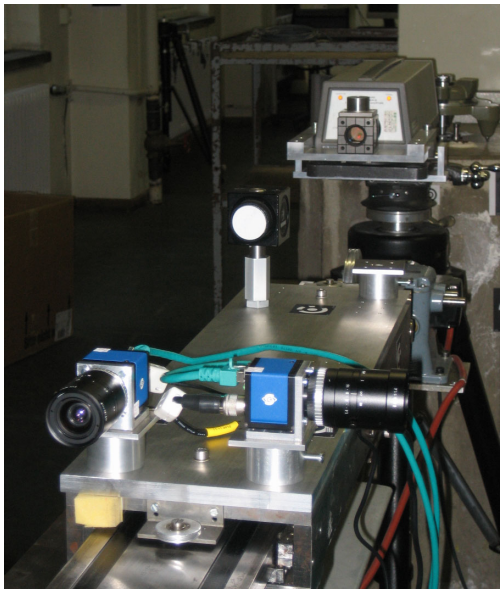


Figure 3. Cameras in orthogonal configuration on platform. The movement is measured by a laser interferometer.

The cameras are directly connected to a PC via a Gigabit-Ethernet adapter using their GigE vision interface. We use two CCTV optics having a different focal length of 6.1 mm and 8 mm respectively with the cameras. The laser interferometer is able to resolve a movement of the platform in linear direction with 1/10 µm. The position's differences can be determined with an accuracy of 2-3 µm. The largest limiting factor is the registration of the atmospheric conditions. Nevertheless for our accuracy requirements it is sufficient that temperature and pressure are determined at one single position. It can be assumed that the environmental conditions are constant for the duration of the experiment. The following Table 4 summarizes the hardware used for the experiment and a schematic layout of the experiment is depicted in Figure 9.

Item	Properties
Camera(s)	SVS-Vistek SVCam eco655 2448 x 2050 pixel 3.45 µm pixelsize monochromatic CCD GigE Vision interface
Optics	VS Technology CCTV 1 SV-0614H, f = 6.1 mm 1 SV-0814H, f = 8 mm
Laser interferometer	HP 5519A

Table 4. Utilized components.

The reference coordinate system is defined by installing circular coded targets on the wall and on several stable auxiliary constructions. Different diameter sizes are used to realize measurements on varying distances. The coordinates for a basic network of 80 coded marks were measured with the Kern ECDS, which resulted in a mean RMS of 0.10 mm and a standard deviation of 0.077 mm. The network is complemented with additional 50 marks for the first of the two scenarios and further 200 different for the second scenario (see section 3.3). These reference points are included into the basic network via images, taken by a photogrammetric camera (NIKON D3).

3.3 Procedure

The experiment is divided into two parts which simulate two different scenarios. The first acquisition situation deals with marks in distances from 2 m up to 9 m (Scenario I) and afterwards an acquisition situation with a larger number of smaller marks within short distances of around 1.2 meters (Scenario II) (see Figure 5) is tested.

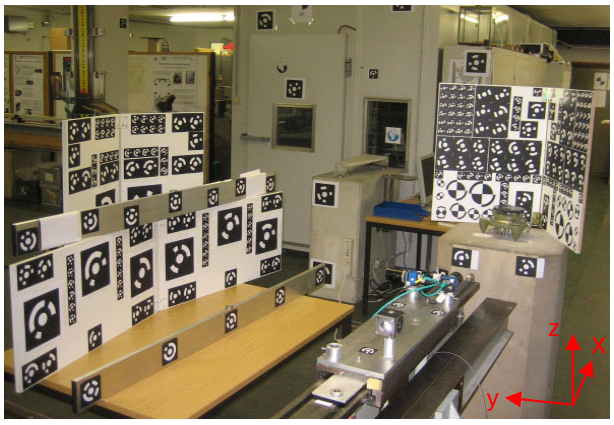


Figure 5. Reference targets in scenario (II)

Different sets of configurations of the camera mounting are carried out. Five configurations are realised in scenario I and six configurations in scenario II. The configurations are depicted in Figure 6 and Figure 7, respectively. Each pair of equally coloured arrows shows one specific configuration for the two used cameras. The shape of the arrowhead denotes the focal length of the optics which was mounted on the camera.

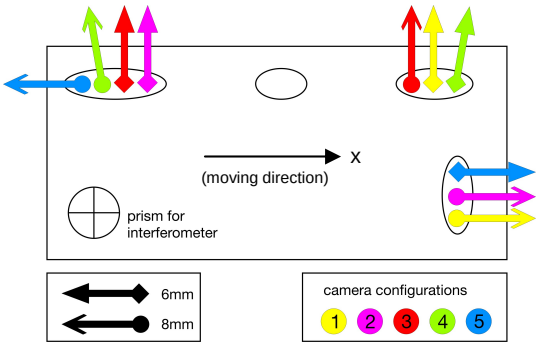


Figure 6. Camera configurations for scenario (I). Each pair of equally coloured arrows denotes a configuration, arrowheads show the used optics.

For scenario I the camera platform is moved to four positions for image acquisition, whereas in scenario II only three positions are taken into account. In Table 8 the mean positions with the related RMS error for all camera configurations are shown. An example for an acquisition situation is depicted in Figure 9.

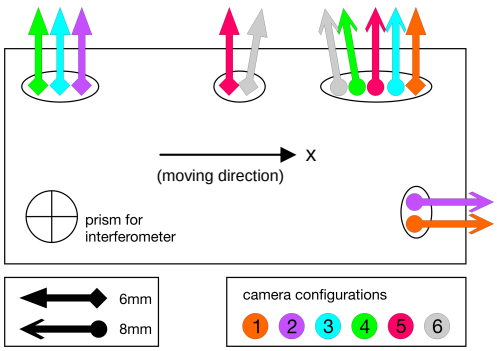


Figure 7. Camera configurations for scenario II. See explanation of Figure 6.

Scenario	Position number	Position on linear axis [mm]	RMS [µm]
I	1	-0.0186	9.6
	2	684.0687	6.7
	3	2437.7995	8.8
	4	2868.9189	3.4
II	1	2711.2399	1.5
	2	2801.4716	2.5
	3	2868.9116	2.5

Table 8. Number of positions and distances for image acquisition.

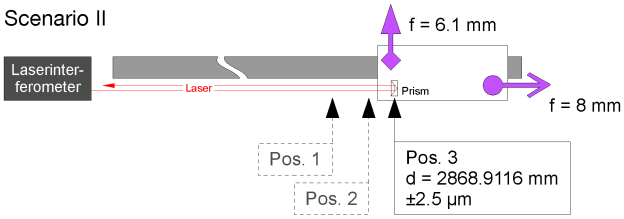


Figure 9. Schematic layout of the experiment and example for image acquisition with configuration 2 at position 3. Values bear on the measurements of the laser interferometer.

4. PROCESSING

The following two analysis strategies are intended for the acquired images: Space resection with a single camera and bundle adjustment using both cameras with overlapping images as well as non-overlapping images.

Furthermore the combination of images taken from cameras in different configurations is planned to simulate a camera setting with more than the actually used two cameras. For this purpose, the positions for image acquisition along the linear slide rail must be met very precisely (see Table 8). The repeatability of a certain position for the carriage, which has to be adjusted manually, is in the range of 10 µm. It must be shown, if this is sufficient for the appropriate merging of the different configurations. Additional variations arise from the fact that cameras with different focal lengths are used.

Since the cameras do not provide an automatic exposure measurement and the illumination situation is expected to be very variable at the test location, an exposure series with constantly increasing exposure time is acquired at every position. This series covers exposure times from 50 ms to 200 ms with an interval of 25 ms.

Data processing will be done in two steps. First the coordinates of all reference targets have to be determined. Incorrectly assigned target codes and outliers have to be corrected or eliminated manually. In a second step, the exterior orientation parameters will be computed within a bundle adjustment exploiting every reasonable combination of the above mentioned variations of image acquisition. In addition occlusions will be simulated by deleting different image measurements from the vector of observations.

The results thus obtained, more precisely the positions of the camera's projection centres, will then be compared to the measurements done with the laser interferometer. For that, coordinate differences between two acquisition positions must be calculated.

Parallel to the data analysis of the experiment a computation derived from a geometrical model of the scene will be done. The aim is to verify the theoretical results with the results from the real test.

5. DISCUSSION

The aim of this first experiment is to obtain 3D coordinates of a rigid camera platform within an absolute coordinate system. For this purpose two cameras with different optics were available only. From a practical point of view this circumstance implies a differing imaging geometry depending on the direction of motion relative to the reference coordinate system (along or across the optical axis), which could be seen as a disadvantage. On the other hand this allows considering a larger number of configurations.

6. FUTURE WORK

A next step would be to extend the experiment in a way that camera orientation angles can be measured and evaluated as well, since according to the experimental setup described in this paper it is only possible to examine a shift.

Further investigations will be made in the field of appropriate reference targets. In cases of partly occluded centres of the circular reference marks, their coordinates cannot be determined correctly anymore. Maybe a reference target could be helpful, that allows the computation of the centre in an additional way. A centre point definition by two crossing lines could be imaginable. An additional advantage would be that the influence of a displaced ellipse centre could be directly measured. This investigation involves, that a decision must be made during the image measurement, to which extraction algorithm a priority is given.

Also a process has to be developed, that a flexible extension of a core reference field is possible. As the first tests have shown, it is often necessary to densify the reference field in some situations. To do this, the user should not be dependant of the availability of another photogrammetric system.

To overcome the problem of occlusions, further test will be made with fisheye lenses. This investigation shall clarify, how far the advantage of a large field of view can compensate the reduction of size of the imaged features. Beyond that a question is how the large distortions at the image margins will allow the recognition of target codes or influence the precision of target measurement at all.

Since at every stop of the camera platform a series of images with different exposure times has been taken, an algorithm has to be implemented, that selects the image measurement of the image, where a certain reference mark is exposed optimal. Furthermore, it needs to be investigated how the additional images can be combined to gain a higher accuracy in the performed image measurements. To obtain a statement for this, the influence of the exposure state on the measured image coordinate needs to be investigated.

ACKNOWLEDGMENTS

We would like to thank the Lehrstuhl für Geodäsie of Technische Universität München for providing us their measurement laboratory and especially Dr. Wasmeier and his team for establishing the basic reference system.
This research is funded by the BFS (Bayerische Forschungsstiftung) within the contract number AZ-876-09.

REFERENCES

Aicon3d, 2011. Product description for the ProCam active probe. http://www.aicon3d.de/produkte/moveinspect-technology/procam/auf-einen-blick.html (accessed 9 May 2011)

van Albada, G.D., Lagerberg, J.M., Visser, A., Hertzberger, L.O., 2007. A low-cost pose-measuring system for robot calibration. *Robotics and Autonomous Systems*, 15(3), pp. 207-227.

Chen, C., Schonfeld, D., 2010. Pose estimation from multiple cameras based on Sylvester's equation. *Computer Vision and Image Understanding*, 114(6), pp. 652-666.

Dold, J., 1996. Influence of Target Size on the Results of Photogrammetric Bundle Adjustment. In: *Proceedings of the ISPRS Commission V Congress*, Vienna, Austria, Vol. XXXI, Part B5, pp. 119-123.

Facchinetti, C.; Tièche, F. & Hügli, H., 1995. Self-Positioning Robot Navigation Using Ceiling Image Sequences. In: *Proceeding ACCV'95 (Asian Conference on Computer Vision)*, Singapore, Vol. 3, pp. 814-818.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), pp. 381-395.

Frahm, J.-M., Köser, K., Koch, R., 2004. Pose Estimation for Multi-camera Systems. In: *Lecture Notes in Computer Science Vol. 3175: Pattern Recognition*, Springer, 2004, pp. 286-293.

Gupta, O. K., Jarvis, R., 2010. Robust pose estimation and tracking system for a mobile robot using a panoramic camera. In: *Proceedings of the 2010 IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, Singapore, pp. 533-539.

Hahn, M., Krüger, L., Wöhler, C., Sagerer, G., Kummert, F., 2010. Spatio-temporal 3D Pose Estimation and Tracking of Human Body Parts in an Industrial Environment. In: *Photogrammetrie - Laserscanning - Optische 3D-Messtechnik: Beiträge der Oldenburger 3D-Tage 2010*, Wichmann, 2010, pp. 158-170.

Hefele, J., Brenner, C., 2001. Robot pose correction using photogrammetric tracking. In: *SPIE Proceedings: Machine Vision and Three-Dimensional Imaging Systems for Inspection and Metrology*, Vol. 4189, pp. 170-178.

Lemaire, T., Berger, C., Jung, I.-K., Lacroix, S., 2007. Vision-Based SLAM: Stereo and Monocular Approaches. *International Journal of Computer Vision*, 74(3), pp. 343-364.

Linkugel, T., Bobey, K., 2010. Visuelle Positions- und Bewegungsbestimmung eines autonomen Systems. In: *Photogrammetrie - Laserscanning - Optische 3D-Messtechnik: Beiträge der Oldenburger 3D-Tage 2010*, Wichmann, 2010, pp. 195-201.

Luhmann, T., 2009. Precision potential of photogrammetric 6DOF pose estimation with a single camera. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(3), pp. 275-284.

Maas, H.-G., 1997. Dynamic photogrammetric calibration of industrial robots. In: *SPIE Proceedings Videometrics V*, San Diego, Vol. 3174, pp. 106-112.

Mautz, R., Tilch, S., 2010. Innenraumpositionierung mit optischen Methoden. *Allgemeine Vermessungs-Nachrichten*, 2010(7), pp. 250-255.

Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F., Sayd, P., 2006. Monocular Vision Based SLAM for Mobile Robots. In: *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*, Hong Kong, Vol. 3, pp. 1027-1031.

Muffert, M., Siegemund, J., Förstner, W., 2010. The estimation of spatial positions by using an omnidirectional camera system. In: *Proceedings of the 2nd International Conference on Machine Control & Guidance)*, Bonn, Germany, pp. 95-104.

Muhle, D., Abraham, S., Heipke, C., Wiggenhagen, M., 2008. Automatische Orientierung von zwei gemeinsam bewegten Stereosystemen ohne gegenseitige Korrespondenzen. In: *Photogrammetrie - Laserscanning - Optische 3D-Messtechnik: Beiträge der Oldenburger 3D-Tage 2008*, Wichmann, 2008, pp. 186-193.

Point Grey, 2011. Product description for the Ladybug3 spherical digital video camera system. http://www.ptgrey.com/products/spherical.asp (accessed 9 May 2011)

Rohrberg, K., 2009. Geschlossene Lösung für den räumlichen Rückwärtsschnitt mit minimalen Objektinformationen. In: *Photogrammetrie - Laserscanning - Optische 3D-Messtechnik: Beiträge der Oldenburger 3D-Tage 2009*, Wichmann, 2009, pp. 332-339.

Sackewitz, M., 2009. Marktstudie 3-D-Messtechnik in der deutschen Automobil- und Zuliefererindustrie. Marktstudie, Fraunhofer-Allianz Vision, Erlangen, Germany.

Sahrhage, V., Riede, R., Luhmann, T., 2006. Optische 3D-Navigation von Sensoren. In: *Photogrammetrie - Laserscanning - Optische 3D-Messtechnik: Beiträge der Oldenburger 3D-Tage 2006*, Wichmann, 2006, pp. 54-61.

Schütze, R.; Boochs, F.; Raab, C.; Wirth, H., Meier, J., 2010. Ein Mehrkamerasystem zur exakten Positionsbestimmung von beweglichen Effektoren. In: *Photogrammetrie - Laserscanning - Optische 3D-Messtechnik: Beiträge der Oldenburger 3D-Tage 2010*, Wichmann, 2010, pp. 220-229.

Wells, G., Venaille, C., Torras, C., 1996. Vision-based robot positioning using neural networks. *Image and Vision Computing*. 14(10), pp. 715-732.

Willert, V., 2010. Bildbasierte Indoor-Positionierung. *Allgemeine Vermessungs-Nachrichten*, 2010(7), pp. 256-263.

MULTI-STEP AND MULTI-PHOTO MATCHING FOR ACCURATE 3D RECONSTRUCTION

M. Previtali, L. Barazzetti, M. Scaioni

Politecnico di Milano, Department of Building Environment Science and Technology
Via M. d'Oggiono, 18/a, 23900 Lecco, Italy
email: mattia.previtali@mail.polimi.it, (luigi.barazzetti, marco.scaioni)@polimi.it,
web: http://www.icet-rilevamento.lecco.polimi.it/

Commission III – WG/1

KEY WORDS: Accuracy, Feature Extraction, Matching, Orientation, Robust Estimation, Surface Measurement

ABSTRACT:

The paper presents an automated procedure for surface reconstruction from digital images. This method was developed for close-range photogrammetric applications, with a particular attention to terrestrial free-form objects that can be modelled with point clouds extracted from images. Therefore, the paper is not directly concerned with architectural elements, where objects feature breaklines and discontinuities that are preferably modelled with manual measurements. The implemented algorithm (MGCM+) integrates two image matching techniques developed in Photogrammetry and Computer Vision in order to obtain metric results in an automated way. Different strategies were exploited to successfully combine both strategies, along with several new improvements. Starting from a set of images and their orientation parameters a preliminary seed model is extracted by using a patch-based algorithm (PMVS). Then, a multi-photo refinement via LSM (MGCM) improves the precision of results and provides a statistical evaluation through a variance-covariance matrix.

1. INTRODUCTION

Nowadays there is an intense research activity aimed at developing new strategies for object reconstruction from images. Several approaches were developed in both Photogrammetry and Computer Vision (CV), keeping in mind different requisites such as accuracy, completeness, automation, reliability, and so on. Furthermore, the typology of the analyzed objects is diverse as well. For instance, in Furukawa et al. (2010), Frahm et al. (2010) and Strecha et al. (2010) some methods for "large scale city modelling" are illustrated, showing impressive results. Several thousands of images are automatically processed in order to obtain 3D models of vast areas containing complex buildings.

Photogrammetrists could state that these kinds of reconstructions are not good for mapping purposes, as results are often incomplete and are not accompanied by statistical analyses. Typical problems can be (i) modelling of architectural objects, where breaklines should be matched in order to obtain sharp edges, (ii) use of uncalibrated cameras and images downloaded from the Internet, without any consideration about network design, (iii) lack of a geodetic network for stability control in the case of large blocks, with GCPs used as pseudo-observations in bundle adjustment.

Accuracy during image orientation becomes a point of primary importance in Photogrammetry. Indeed, a photogrammetric bundle adjustment is supposed to ensure the metric quality of the final result. This is partially in contrast with a CV bundle adjustment, as best summed up in Snavely et al. (2008), where the functioning of Bundler is described: "most SfM methods operate by minimizing reprojection error and do not provide guarantees on metric accuracy".

This different point of view is motivated by the use of the final 3D model. In a few words, the approaches are diverse because purposes are diverse.

However, in close-range photogrammetry a growing number of CV methods is receiving great attention. For the image orientation phase, it is now available on the market the new PhotoModeler 2011 (EOS, Canada – www.photomodeler.com), that is the first (photogrammetric) commercial package capable of orienting target-less images in a fully automated way. The mathematical model used during bundle adjustment is a typical photogrammetric approach, but the operator for image matching (SIFT in this case) and the strategies for outlier rejection (based on the *fundamental* or *essential matrices*) come from CV. In addition, there are also other solutions for automatic orientation in close-range, where different techniques (e.g. *Least Squares Matching* - LSM) are integrated to improve precision and reliability (Barazzetti et al., 2010; Pierrot Deseilligny and Clery, 2011; Roncella et al., 2011).

This means that the combined use of techniques developed in both disciplines allows one to obtain accurate results in a fully automated way. This is now a reality for image orientation only, while with the work presented in this paper we would like to extend the concept also towards 3D modelling (for some specific categories of objects).

In the field of close-range photogrammetry some (multi-image) commercial software for surface reconstruction are available today. Most of them are derived from Aerial Photogrammetry (e.g. CLORAMA - Remondino et al., 2008; LPS eATE - www.erdas.com). In their original implementations, these are able to extract a digital surface model (DSM), that is a 2.5D representation of the ground. On the other hand such 2.5D models are adequate for airborne mapping, but they feature evident limits in close-range surveys, because they cannot handle scenes at 360°. In addition, problems arise for DSM cells having multiple depth values. On the other hand, the accuracy obtainable with these methods is noteworthy, especially thanks to the use of sub-pixel *area-based matching* (ABM) procedures.

As previously mentioned, we would like to present a methodology for surface measurement in the case of 3D objects. The aim is to obtain models useful for photogrammetric surveys. Architectural objects with sharp breaklines (e.g.

facades) are not considered here because their detailed modelling can best be accomplished with interactive measurements, through the identification of basic geometric shapes. We focus instead on 'free-form' objects, i.e. objects that can be modelled with meshes generated from point clouds.
The proposed matching procedure is divided into two steps. First of all a seed model is created with a *patch-based image matching* technique and then ABM operators densify and refine the point cloud. Multiple images are simultaneously used to detect and remove outliers with the analysis of the light ray intersections in 3D space. As the core for global processing is the combined use of *Multi-photo Geometrically Constrained Matching* (MGCM) and other methods (that essentially allow 3D processing), we called the procedure MGCM+.

2. 3D RECONSTRUCTION PIPELINE WITH MGCM+

In this section a complete pipeline for the geometrical reconstruction of large-scale objects and scenes, using high resolution images in a multi-view stereo framework, is described. This method incorporates the high-end matching algorithms developed in Photogrammetry and CV.
As can be seen in the flowchart in Figure 1, a block of suitable images (in terms of network geometry and image resolution) is needed. All images must be captured by using calibrated cameras in order to improve the precision of the final 3D measurements. It is out of the scope of this paper to illustrate the geometric characteristics that the image block should have (overlap, external or internal constraints, relative angles between images and the like). To obtain accurate and reliable surface measurements, each portion of the whole object must be covered by at least 3-4 images to exploit the potential of multi-photo matching. In addition, the length of any baseline has to be selected according to a compromise between the precision in the depth direction (large baselines are better) and the limitation of image deformations required by ABM (short baselines and small view angles).
Image orientation is the second prerequisite of MGCM+. It can be performed by manual or automatic procedures, but the latter have the advantage to generate a denser point cloud of tie points. This can be used to initialise MGCM+ with an approximate surface. On the contrary, if this initial model is not sufficiently dense, an alternative solution is applied to derive the seed model. This is mainly based on the algorithm proposed by Furukawa and Ponce (2010), as illustrated in subsection 2.1. The advantage of this method is its independence from the reference frame adopted and the capability of working without any initial rough model of the object.
The MGCM algorithm combines (i) LSM based on intensity observations with (ii) collinearity conditions used as geometrical constraints for the determination of all object point coordinates. The introduction of the collinearity constraints and the opportunity to simultaneously match multiple scenes increase the matching reliability.
MGCM is a matching technique presented by Grün (1985), with new improvements in Grün and Baltsavias (1988) and Baltsavias (1991). One could say that this technique is old. However, it is still the most precise method for image coordinate measurement and it is therefore appropriate for metric purposes. Here the theoretical background of MGCM is not explained in detail, but the main aspects of this technique are outlined to highlight its advantages with respect to other ABM methods. In addition, some limitations of the original formulation are pointed out.

Let us consider a block of images depicting an object. One of them is selected as 'master image' according to a specific criterion (see subsect. 2.2). A set of points (x_{Pk},y_{Pk}) found by means of an *interest operator* (or nodes of a regular grid) are used to extract a set of 'templates' i.e. square patches with a side of a few pixels. Each of them is reprojected on the other images of the block by exploiting a rough DSM of the object. A squared window ('slave') is extracted around each reprojected point for each generic image i, obtaining a total number of n possible candidates.

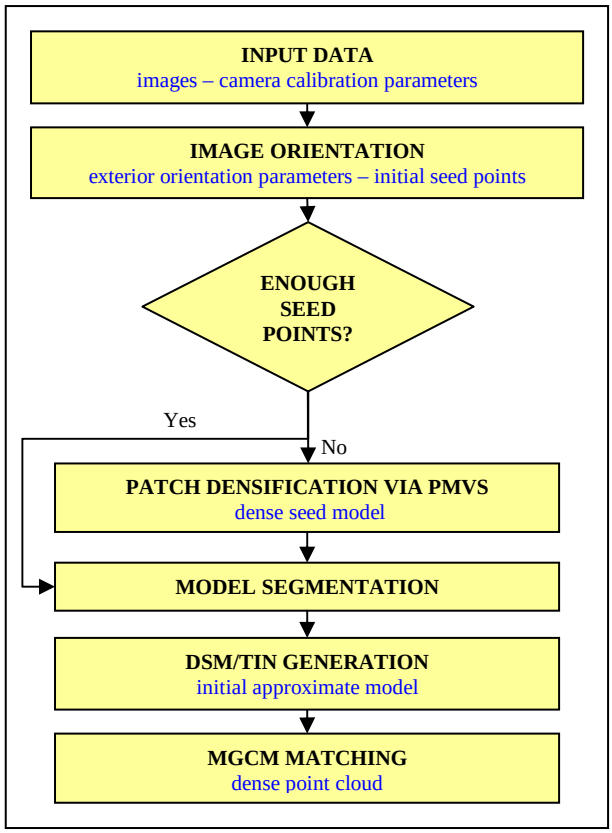


Figure 1. Workflow of MGCM+.

The geometric deformation between the 'template' and each 'slave' is modelled using an affine transformation, which locally approximates quite well perspective deformations. Then the 'template' is compared with all corresponding 'slaves'. The relationship describing the intensity values of each pixel in the 'template' is given by the discrete function $f(x,y)$, and the n 'slaves' are represented by functions $g_1(x,y)$, …, $g_n(x,y)$. An intensity observation equation for each pixel of the 'template' and the corresponding pixel on the 'slave' i is written as follows:

$$f(x,y) - e_i(x,y) = g_i^0(x,y) + g_{xi}\,da_{10\,i} + g_{xi}\,x_o\,da_{11\,i} + g_{xi}\,y_o\,da_{12\,i} + g_{yi}\,da_{20\,i} + g_{yi}\,x_o\,da_{21\,i} + g_{yi}\,y_o\,da_{22\,i} \quad (1)$$

where the unknown quantities are corrections to the parameters of the affine transformation da_{jki}. The coefficient $g_i^0(x,y)$ is the observed value in the approximate position of the 'slave', while g_{xi} and g_{yi} are the partial derivatives of the function $g(x,y)$. Numerically, the derivatives correspond to row and column

gradients. The function $e_i(x,y)$ gives the residual error with respect to the affine model.

In addition, it is also possible to take into account some radiometric transformations (in many cases using a linear formulation) between 'template' and 'slave'. However, in our case we usually disregard this radiometric compensation to limit the parameters and we prefer to operate with a preliminary radiometric equalization at local level.

The MGCM combines the intensity observation equations (1) with the collinearity condition. In fact, for a pinhole (central perspective) image k the constraint between the generic object point ($X_p=[X_p\ Y_p\ Z_p]^T$) and its corresponding 2D point (x_{pk},y_{pk}) on the image k is given by the well-known collinearity equations:

$$x_{pk} = -c_k \frac{r_{1k}^T \left(X_p - X_{ok} \right)}{r_{3k}^T \left(X_p - X_{ok} \right)} \hat{=} -F_k^x \qquad (2)$$

$$y_{pk} = -c_k \frac{r_{2k}^T \left(X_p - X_{ok} \right)}{r_{3k}^T \left(X_p - X_{ok} \right)} \hat{=} -F_k^y$$

where c_k is the principal distance, X_{0k} is the vector expressing the perspective centre coordinates, $\mathbf{R}_k=[r_{1k}\ r_{2k}\ r_{3k}]^T$ is the rotation matrix. Image coordinates (x_{pk},y_{pk}) are computed with respect to the principal point.

If both interior and exterior orientation (EO) parameters of each station are known, eq.s 2 can be rewritten as follows:

$$\begin{aligned} \Delta x_{pk} + F_k^x + x_{pk}^0 = 0 \\ \Delta y_{pk} + F_k^y + y_{pk}^0 = 0 \end{aligned} \qquad (3)$$

The unknown parameters in eq.s 3 are shifts (Δx_k, Δy_k) and object point coordinates (X_p). After their linearization, Eq.s 3 become:

$$\begin{aligned} \Delta x_{pk} + \frac{\partial F_k^x}{\partial X} dX + \frac{\partial F_k^x}{\partial Y} dY + \frac{\partial F_k^x}{\partial Z} dZ + F_k^{x(0)} + x_{pk}^0 = 0 \\ \Delta y_{pk} + \frac{\partial F_k^y}{\partial X} dX + \frac{\partial F_k^y}{\partial Y} dY + \frac{\partial F_k^y}{\partial Z} dZ + F_k^{y(0)} + y_{pk}^0 = 0 \end{aligned} \qquad (4)$$

Shifts allow one to link both sets of eq.s 1 and 4, because $\Delta x_{pk}=da_{10}$ and $\Delta y_{pk}=da_{20}$ for the same set of images and point P. Therefore, the resulting joint system can be solved using conventional Least Squares solution schemes (see Baltsavias, 1991).

MGCM presents some important advantages with respect to other traditional automatic matching techniques used in Photogrammetry and CV. Compared to the normal LSM, where it is possible to match simultaneously only a couple of images, the MGCM obtains highly redundant results thanks to the collinearity constraint that permits combined multi-image matching. This reduces multiple solutions in case of repetitive textures and helps overcome possible occlusions thanks to the chance to view the object from multiple stations. In addition, 3D object point coordinates are directly computed together with their theoretical accuracies. Recently, an extensions of the standard cross-correlation technique have been developed, obtaining the so called *Geometrically Constrained Cross-Correlation* (GC^3) (Zhang and Grün, 2006). This technique uses the collinearity condition in a way similar to MGCM. However, the perspective changes in close-range data can cause some troubles to the correlation strategy, while the affine transformation between the template and each slave increases the potential of MGCM algorithm in case of convergent images.

The application of our method requires a decomposition of the object into 2.5D regions. For each of them a DSM or a TIN (triangulated irregular network) structure are interpolated starting from a set of seed points (subsect. 2.1). In particular, the choice of the subset of images for the measurement of each portion of the object is a crucial task. Inside this problem, a key aspect is which image could better serve as 'template' (subsect. 2.2). All portions reconstructed in the local reference systems are finally joint together to derive a unique 3D model. It is important to mention that the recombination of the point clouds is rigorous as the rigid body transformations employed are known exactly. With this approach, even though the core matching strategy is still 2.5D, any 3D shape could be potentially reconstructed.

2.1 Seed model generation

As the measurement of the object surface with the MGCM algorithm needs an initial approximation, an intermediate step was added to obtain a preliminary seed model. In the case the EO parameters have been computed by using an automatic procedure, one could use all tie points matched with feature-based matching (FBM) operators. However, in some cases their number is not sufficient (e.g. with texture-less objects) or their distribution in the images can be really variable, leaving some empty areas.

The importance of a good seed model is remarkable not only for the geometric quality of the final product, but also in terms of CPU time as it can limit the search along the 3D light ray, reducing the number of trials during the translation of the correlation window. Lastly, tie-point coordinates are usually incorporated into a photogrammetric bundle adjustment. If the number of point correspondences used for image orientation becomes significant, there is a consequent increment of the computational cost. According to the authors' experience some (few) close-range photogrammetric packages can process several thousand of image points, but when there are more than 100,000 image coordinates, the computation of a rigorous bundle solution based on collinearity equations could become quite difficult, especially with standard PCs.

For this reason a limited number of tie points with a good distribution and geometric multiplicity is still the best compromise during the orientation phase. The generation of a seed model is carried out in a new matching phase, where EO parameters are kept fixed in order to exploit the geometric constraint due to collinearity.

As our method was developed for 3D objects, we exploit the *patch-based matching* (PMVS) approach proposed by Furukawa and Ponce (2010). Their procedure was incorporated into our matching pipeline in order to generate a low resolution initial model. This choice is motivated by the robustness of the method that combines multiple images during the dense matching step: if at least three images are processed simultaneously, blunders and spurious points can be removed by analysing the local data redundancy. In addition, the method is able to work with 3D objects and does not require any manual measurement.

With these considerations in mind, the use of an intermediate procedure in the reconstruction pipeline could be seen as a drawback and a lack of originality. On the other hand, it is quite difficult to find (or develop) an open source implementation much better than Furukawa and Ponce's code. In addition, as mentioned in the introduction, this work is partially based on

techniques developed by different authors. Nowadays, some of these have reached a significant level of maturity, while others still need improvements to become useful in practise. Our contribution tries to combine different procedures and this is not a trivial task. Moreover the processing pipeline should be able to exploit only the best properties of each technique. For instance, there are two fundamental drawbacks found with the implementation available:

- it is difficult to manage several high resolution images used at their original size; and
- the input data are expressed with the P-matrix camera model (Hartley and Zisserman, 2003), while a photogrammetric bundle adjustment provides EO parameters.

The solution for the second point can be found in Barazzetti (2010), where an explicit relationship between both orientation datasets is illustrated for the case of distortion-free images. The first problem can be overcome by using compressed images with an opportune modification of the projection matrices. This solution is viable because we are interested in the creation of a seed model. In several scientific work images are subsampled during dense matching without introducing a coarse-to-fine approach that considers the original data. This is an approximation that degrades the quality of the final result, and it assumes an increasing importance due to the technological improvement of digital cameras, with geometric resolutions superior to 14 Mpix even for low-cost compact sensors. Our solution to this problem is described in the next section, where images are always used at their original size without any loss of precision.

2.2 Approximate geometrical model handling

An important limit in the MGCM algorithm is the need of an approximate position of the object points, along with a preliminary location of the homologous point positions in the slave images. To overcome this problem a seed model of the object is derived by using the method described in the previous section. In the current implementation of MGCM+ the object surface is approximated by using a DSM oriented with respect to a reference plane. In the case of complex 3D objects which do not feature a 2.5D geometry (likewise the topographic surface in mapping projects), the whole patch-based point cloud is segmented in approximated 2.5D regions. As things stand now, the segmentation of the point cloud is performed in a manual way (this solution is fast and simple for many objects). Each model is processed in a separate way and the final point clouds are then connected together. This is also good for parallel computing. In any case, the use of a more flexible 3D data structure like a TIN will be added soon, because it is the best solution for complex objects.
A set of points is then defined in the 2D regular grid of each DSM. The cell size can be set by the user according to the resolution of the images. The approximate elevation of each point with respect to the reference plane is selected by using interpolation techniques. In a second stage, this elevation will be estimated by the MGCM L.S. solution. A back-projection of each grid point on all the available views is then carried out to select the 'master' and 'slaves' images and it also provides the set of initial positions of homologous points.
However, although the DSM used can be a rough approximation of the real surface, the homologous points defined by the back-projection principle can be very far from their true positions. For this reason, additional intermediate points are set up on the projective ray connecting the DSM cell and the image point in the 'master' image. The number of additional points along the projective ray and their inter-distances are both parameters estimable on the basis of the approximated surface model quality. For each point defined along the projective ray the 'slave' image patches are derived using the collinearity principle. According to this approach, a set of approximated candidate positions for the L.S. solution of the system is found. At this stage, both sets of eq.s of kind (1) and (4) are set up in order to compute the corrections for image and object point positions. The partial variance factors σ_{0i}^2 for each individual patch are also estimated. The process is iterated until the corrections are negligible. The last problem concerns the choice, among all candidate solutions computed along the projective ray, of the correct matching. In particular, it is considered as correct match the one minimizing the mean variance factor.

2.3 Selection of images and LSM approximate parameters

As remarked above, one of the weak points in the original MGCM formulation is the selection of 'master' and 'slave' images. Generally the problem is solved in this way: given the set of images to be processed, an image (usually the central one) is manually picked up as 'master'. Consequently, all the other images will serve as 'slaves'. The manual choice of a fixed 'master', obviously, is not the best criterion. This is mainly due to a couple of reasons: (i) if an object, approximated with a 2.5D model, is not entirely visible in a single image, multiple processing with different 'master' images is needed; (ii) in terrestrial surveys there are some lateral views of an object, capturing areas occluded in the central image. They give an important contribution to the final reconstruction of the object. This contribution would be completely neglected by using a 'fixed master' approach.
Nevertheless, the alternate use of all photos as 'template' is not a good solution because of the huge CPU time needed to complete global processing. Finally, in close-range applications the perspective deformations between different images can be so large that the affine model between 'master' and 'slave' images could become inadequate.
For all these reasons an optimization in the image selection phase is needed. To start with, a selection based on the information derived from the approximated model is accomplished. For a specified point in the DSM, all images in which the point is visible are considered with a simple back-projection. The selection of the 'master' is then carried out inside this set. The surface normal direction in correspondence to the considered object point is computed, then this direction is compared to all photo normals where the point is visible. The image whose normal is closer to the surface normal direction is chosen as 'master'. With this strategy we can easily handle also 2.5D objects that are not completely visible in a single image, without requiring the intermediate interaction of the user.
An optimization is also mandatory for what concerns the selection of 'slaves' images. In fact, in many cases the perspective deformations can become a problem for LSM. For this reason we limit the number of possible 'slaves' only to those where LSM can provide good results. Also in this case the choice is operated using the approximated surface model. For each point in the original model we consider the shape of the DSM cell containing the same points in different images. If a large geometrical deformation occurs, the shape of the DSM cell in the images presents significant changes. Therefore we back-project the DSM cell containing the object points in all images and we compare the cell changes between the defined

'master' image and the other 'slave' candidates. In particular we consider two parameters: 'cell area' and 'cell shape'. If the area of the back-projected DSM cell in a 'slave' is less than half of the cell area in the 'master', the variation between images, owing to both perspective changes or scale variations, is considered too large and the point is not processed. However, in some cases even if the area does not vary too much from two images a significant perspective variation could occur. In this case the "shape" of the cell changes in a significant way. To recognize this situation we consider the inclination of the back-projected DSM cell on the images. If angular variations between the 'template' and a candidate 'slave' are superior to 40% the image is rejected.

Finally, it is important to find a set of approximate parameters for the affine transformation between 'template' and 'slaves'. After selecting the 'slave' images as described above, the DSM cell is known in different images. This information can be used to compute initial values for rotation, affinity and scale parameters for the LSM, simply using an affine transformation between the back-projected DSM cell in the 'master' and 'slave' images. As shown in Balsavias, (1991) the significance of the shaping parameters in the affine transformation can be evaluated with their correlations. In fact, high correlations among the parameters of the affine model and the others might indicate their non-determinability. In our case we are particularly interested in evaluating the significance of shears and scales as approximate values. The correlations between similar shaping parameters (scales - shears) and the correlations between shape parameters in the same direction have an essential importance. At this stage two approaches can be used to evaluate these correlations: a deterministic approach and a statistical one.

In the first case all parameters can be considered as highly correlated if their correlation coefficients exceed a fixed threshold. This means that one of the two correlated parameters can be assumed as not significant and the one with the larger variance should be excluded.

In many cases the use of a fixed threshold for the definition of high correlations can be a real challenge and could lead to a poor solution. Here, a statistical approach becomes more suitable. In particular, it is possible to assume that the parameters have a multivariate normal density distribution. Under this hypothesis, and after fixing a significance level for the test, the correlation of the shaping parameters can be verified using *Hotelling's test* (Baltsavias, 1991). If the test fails, there are correlations, otherwise all parameters can be considered as statistically uncorrelated. A further investigation should be carried out to determine which coefficients are effectively correlated. This check can be done with a test that imposes the null hypothesis $\rho = 0$.

As can be noticed, in the statistical approach no threshold needs to be set at the beginning. It is necessary to fix only the significance level for the test. Therefore, this is the default procedure in our method.

3. EXPERIMENTS

Shown in figures 2a-b-c are three close-range objects modelled from multiple convergent images (12 Mpix) by using the proposed method. The first example (a) is a 3D object that was divided into three portions to fit the 2.5D requisite. The approximate DSM (its first region is shown with a colour-map representation) was the starting point to extract 1.3 million points roughly. No blunders were found at the end of matching phase and a final mesh was created after the combination of all point clouds. As previously mentioned, the alignment of partial reconstructions does not introduce new errors, as each rigid-body transformation is known.

In this case a rigorous accuracy analysis was impossible, as a reference dataset was not available. However, this method allows a statistical evaluation with the covariance matrix, offering the standard deviations of all 3D points.

The second dataset (b) can be clearly modelled without any partitioning. The point cloud obtained from 5 convergent images is made up of 0.5 million points. Blunders were correctly removed by MGCM+. For this dataset we carried out a comparison between another point cloud generated using Leica Photogrammetry Suite (LPS) - eATE. This software was developed for aerial mapping purposes and can be considered a well-assessed tool for object reconstruction, based on *semi-global matching* (Hirschmüller, 2008). Both meshes were aligned using the ICP registration algorithm implemented in Geomagic Studio. The range of the error bar is ±13 mm, while the standard deviation of discrepancies is ±0.7 mm. The object is 0.9 m wide.

The example in Figure 2c comprehends 7 images capturing a bas-relief 1 m wide. Also in this case the object can be easily modelled using a single partitioning, offering the opportunity for a new comparison with LPS-eATE. This gave a discrepancy of about 0.4 mm, while the error bar ranges from [-13; +13] mm.

Some other datasets for multi-view stereo evaluation are available on the Internet (provided by Strecha). They are quite challenging because of a repetitive texture and several breaklines. These are typical objects for which usually a manual reconstruction gives better results. The façade of the building in Figure 2d was modelled using 25 images, that were oriented in order to obtain photogrammetric EO parameters (although the camera-matrices are available). The output of the multi-image matching phase with an incorporated block sub-division are 2.7 million points. For the second dataset (Figure 2e), 11 images were employed to obtain 1.2 million points that were interpolated to create a mesh. A scale factor was then fixed to remove this ambiguity, and the model was aligned with the laser mesh with ICP obtaining a discrepancy (in terms of standard deviation) of about ±12 mm. In any case, during this comparison we included all areas that were not visible in the images, where there are evident gaps in the photogrammetric model. Here, the distances between the model (that are not errors but only empty areas) are superior to 10 cm, and caused a global worsening. The analysis was repeated only for small portions of the model in order to avoid this problem, estimating a std.dev. of about ±5.6 mm. The obtained value is comparable to the ground sample distance (GSD) of the images and the sampling step during MGCM+.

Figure 2f shows a 360° reconstruction from a set of 32 images around a statue, confirming the suitability of the method for 3D objects. The small objects (g) is instead made of marble. Although it is well-known that this material is prone to produce noisy results, the visual reconstruction seems good.

Figure 2h shows a geotechnical case, where a rock face was surveyed at different epochs in order to monitor its stability and discover potential risks (e.g. rockfalls). To accomplish this task the metric content of the model is essential. In addition, as data must be compared to obtain a multi-temporal analysis, images must be registered to the same reference system with some GCPs incorporated into the bundle adjustment. The comparison with a laser model, after removing disturbing elements such as vegetation, revealed a discrepancy of about ±5 mm, i.e. the nominal precision of the laser scanner used.

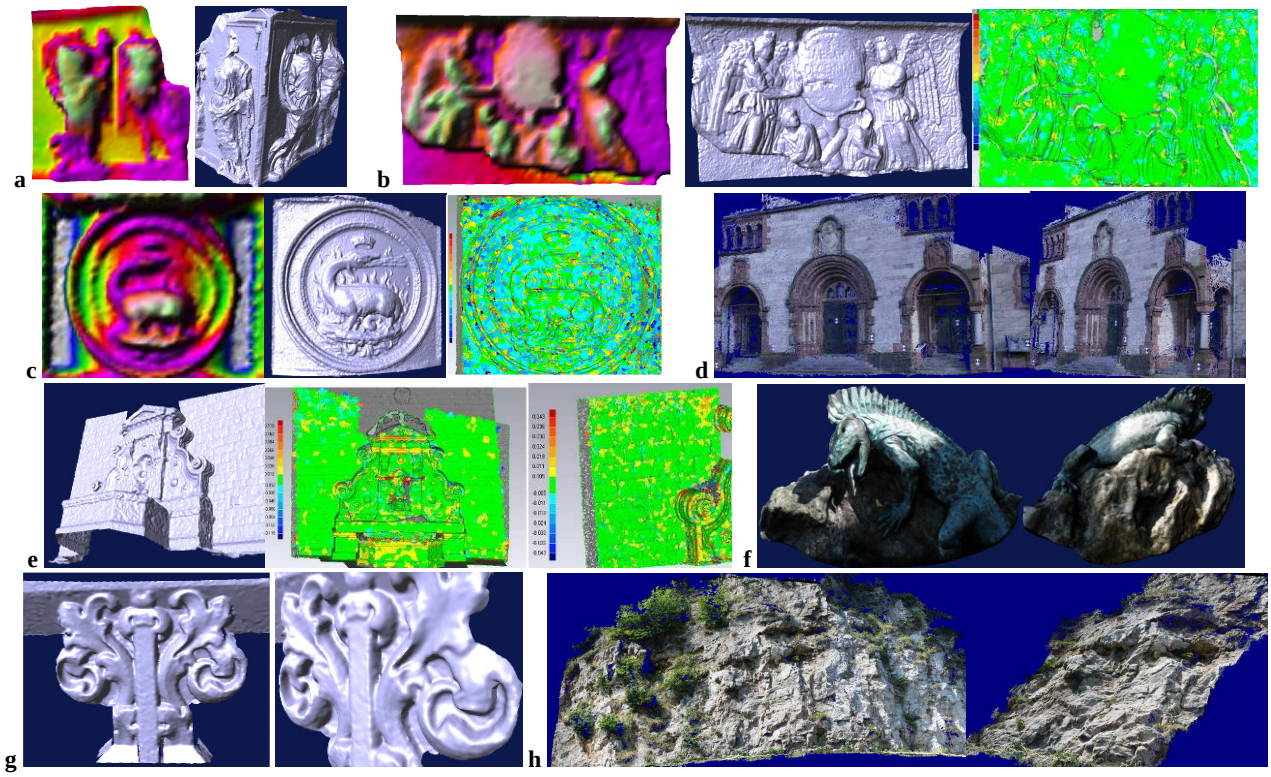


Figure 2. Some results obtained with MPGC+ for the reconstruction of 2.5D and 3D objects.

4. CONCLUSIONS

The paper presented an automated pipeline for multi-view reconstruction of close-range objects. The final aim was to setup a software able to model free-form objects from images featuring good characteristics in terms of resolution, overlap and network geometry. The reconstruction process is automated, starting from image orientation phase up to the generation of a dense point cloud. Partitioning of the object is the only manual task, although an automated solution is under development. Here, we are not interested in the final step, i.e. mesh generation, as several commercial and open source solutions are available to accomplish this task.

An important aspect of this work is the joint use of CV and Photogrammetry techniques. In particular, we think that MGCM is a powerful matching method, as it is very robust, precise and invariant with respect to affine deformations (after setting good initial values). On the contrary, other CV methods can automate the typical photogrammetric workflow. In a few words, the advantages of both disciplines are combined, while shortcomings are reduced.

There are some limits in our approach, like the use of a 2.5D DSM as initial model. A solution based on a 3D TIN as seed model is under investigation. Another limit is the manual segmentation of complex objects. In our future work we will try to eliminate this step or, at least, to introduce an automatic segmentation procedure.

REFERENCES

Baltsavias, E.P, 1991. Multiphoto Geometrically Constrained Matching. Ph. D. thesis, Inst. of Geodesy and Photogrammetry, ETH Zurich, Switzerland, Mitteilungen No. 49, 221 pp.

Barazzetti, L., 2010. A trip to Rome: physical replicas of historical objects created in a fully automated way from photos. Proc. of HCITOCH. *Lecture Notes in Computer Science*, 6529: 63-72 (2011).

Barazzetti L., Remondino, F. and Scaioni M., 2010. Extraction of accurate tie points for automated pose estimation of close-range blocks. *ISPRS Technical Commission III Symposium on Photogrammetric Computer Vision and Image Analysis* , 6 pp.

Frahm, J. et al. 2010. Building Rome on a cloudless day. In: Proc. of ECCV 2010, 14 pp.

Furukawa, Y. and Ponce, J., 2010. Accurate, dense, and robust multi-view stereopsis. *IEEE Trans. PAMI*, 32(8): 1362-1376.

Furukawa, Y., Curless, B., Seitz, S.M. and Szeliski, R., 2010. Towards Internet-scale Multi-view Stereo. Proc. of IEEE Conf. CVPR'10, 8 pp.

Hartley, R., and A. Zisserman, 2003. *Multiple View Geometry in Computer Vision – 2nd Ed.* Cambridge University Press, UK.

Hirschmüller, H., 2008. Stereo processing by Semi-Global Matching and Mutual Information. *IEEE Trans. PAMI*, 30(2): 328–341.

Grün, A., 1985. Adaptive least square correlation – a powerful image matching technique. *S. Afr. J. of. Photogrammetry, Remote Sensing and Cartography*, 14(3), pp 175-187.

Grün, A. and Baltsavias, E.P. 1988. Geometrically Constrained Multiphoto Matching. *PE&RS*, 54(5), pp. 663-671.

Pierrot Deseilligny, M. and Clery, I., 2011. APERO, an open source bundle adjustment software for automatic calibration and orientation of set of images. *IAPRS&SIS*, 38(5/W16), on CD-ROM.

Remondino, F., El-Hakim, S.F., Grün, A. and Zhang, L., 2008. Turning images into 3-D models - development and performance analysis of image matching for detailed surface reconstruction of heritage objects. *IEEE Signal Processing Magazine*, 25(4), pp. 55-65.

Roncella, R., Re, C. and Forlani, G., 2011. Performance evaluation of a structure and motion strategy in architecture and cultural heritage. *IAPRS&SIS*, 38(5/W16), on CD-ROM.

Snavely, N., Seitz, S.M. and Szeliski, R., 2008. Modeling the world from internet photo collections. *IJCV*, 80(2), pp. 189-210.

Strecha, C.,Pylvanainen, T. and Fua, P., 2010. Dynamic and Scalable Large Scale Image Reconstruction. Proc. of CVPR'10, 8 pp.

Zhang, L., and Grün, A., 2006. Multi-image matching for DSM generation from IKONOS imagery. ISPRS Journal, 60, pp. 195-211.

AREA BASED STEREO IMAGE MATCHING TECHNIQUE USING HAUSDORFF DISTANCE AND TEXTURE ANALYSIS

Jyoti Joglekar [a, *], Shirish S. Gedam [b]

[a] CSRE, IIT Bombay, Doctoral Student, Mumbai, India – jyotij@iitb.ac.in
[b] Centre of Studies in Resources Engineering, IIT Bombay, Associate Professor, Mumbai, India – shirish@iitb.ac.in

Working Group III/4

KEY WORDS: Area, Image, Matching, Texture, Hausdorff distance, Reconstruction

ABSTRACT:

A conventional image matching techniques may be classified as either area based or feature based methods. In this paper an area based image matching method is proposed for dense disparity map. The method is a composite technique where first the similarity measure between template window and search window is found by normalized cross correlation technique. Few best matches are selected for the template window from the search sub windows, considering the largest normalized cross correlation coefficient. Further edge map is obtained for stereo image pair using canny edge detector. The matches for the template window are filtered using Hausdorff distance technique. Further texture analysis of the same template window and selected search windows is the third measure to decide the accurate match. Texture analysis is done with the co-occurrence matrices which is a two dimensional histogram of the occurrence of pair of intensity value in a given spatial relationship. With this composite method dense point to point correspondence can be achieved with greater accuracy. This method is tolerant to radiometric distortions and parallel processing of the three techniques will improve the speed.

1. INTRODUCTION

Stereo Image matching is one of the core research areas in Computer Vision and digital Photogrammetry. Technological developments in stereo image matching have advanced from the primitive area based cross correlation technique to more and more precise feature based matching. Stereo allows us to recover information from the given two images about a three dimensional location of objects, which does not exist in any single image. The main goal of stereo image matching is to recover depth information from the given two or multiple images (Zitova and Flusser, 2003).

In order to recover depth information the stereo images should be brought into point-point correspondence. Correspondence points are the projections of a single point in to the three-dimensional scene. The difference between the locations of these two correspondence points is known as parallax or disparity which is a function of position of the point in the scene, orientation and physical characteristics of the camera. So epipolar line and disparity can be used as constraints for matching. Though feature based technique is more accurate, it gives sparse disparity map. We are proposing an area based method as it gives dense disparity map for 3D reconstruction.

The techniques used are zero mean normalized cross correlation, Hausdorff distance and texture analysis. In this method the basic units that are used for matching is regularly sized neighbourhood of a pixel. The position of the given pattern is determined by a pixel wise comparison of the image with a given template that contains the desired pattern. For this the template is shifted m discrete steps in x direction and n discrete steps in the y direction of the image and the comparison is calculated over the template area for each position (m, n).

Dense depth measurements are required in applications such as teleconferencing, robot navigation and control. As the taxonomy given in (Peleg, and Weiser, 1996), stereo algorithms that generate dense depth measurements can be roughly divided in to two classes, namely global and local algorithms. Global algorithms (Fusiello, Roberto, and Trucco, 2000) based on iterative schemes that carry out disparity assignments on the basis of the minimization of a global cost function.

As compared to local algorithm, global algorithm yield accurate and dense disparity measurement but computational cost of these algorithms is not suitable for real time application.

Local algorithms (Muhlmann, Maier, Hesser and Manner, 2002; Fusiello, Trucco and Verri, 2000; Trucco and Verri,1998) also referred as area-based algorithms calculate the disparity at each pixel based on photometric properties of the neighboring pixels compared to global algorithm, local algorithms yield significantly less accurate disparity maps. So to improve the accuracy the composite algorithm is proposed here. By selecting the disparity range found out by visual inspection of stereo image pair the search area can be limited and the computational speed can be improved so that the matching algorithm can be used for real time applications.

Area based method uses pixel intensity directly to compute the similarity measure between a small template and a large search window by the quantities of statistical correlation, Fourier cross power spectrum or mutual information. The area based methods merge the feature detection step with the matching part. These methods deal with the images without attempting to detect salient objects. There is a high probability that a window containing a smooth area without any prominent details will be matched incorrectly with other smooth areas in the reference image due to its saliency.

Classical area based methods like cross correlation exploit for matching directly image intensities without any structural analysis. Consequently they are sensitive to the intensity changes, introduced by noise, varying illumination and/or by using different sensor types.

There are generalized versions of cc for geometrically more deformed images. They compute the cc for each assumed geometric transformation of the sensed image window. Similar to the cc methods is the sequential similarity detection algorithm SSPA (Egnal and Wildes, 2000). It uses the sequential search approach and a computationally simpler distance measure than the cc. it accumulates the sum of absolute differences of the image intensity values. The method is likely to be less accurate than the cc but it is faster.

Two main drawbacks of the correlation like methods are the flatness similarity measure maxima and high computational complexity. In the proposed algorithm reducing the search window by finding disparity range with visual inspection the search area is limited. Hence the computational complexity is reduced. Further structural analysis of the probable matches is done with Hausdorff distance criteria and texture analysis to improve accuracy of the match.

The Section 2 explains three techniques used in the proposed algorithm. Section 3 explains the proposed algorithm, which is a composite technique using zero mean normal cross correlation, Hausdorff distance and texture analysis. In section 4, results are presented for a sample stereo image pair. Finally the analysis of the result and an outlook to future research activities is presented.

2. TECHNIQUES USED FOR IMAGE MATCHING

2.1 Zero mean normalized cross correlation algorithm

The problem treated in this paper is to determine the position of a given pattern in a two dimensional image f. Let f(x, y) denote the intensity value of the image f of size M x N at the point (x, y), x Є { 0, 1, ……, M-1}, y Є {0,1, ……,N-1}. The pattern is represented by a given template t of size (p x q). A common way to calculate the position (m_{pos}, n_{pos}) of the pattern in the image f is to evaluate the zero mean normal cross correlation value ρ, at each position (m, n) for f and template t which has been shifted m steps in x direction and n steps in y direction. The size of search window which is larger than template window is determined by visual inspection of the stereo image pair with horizontal and vertical disparity.

Equation (1) gives a basic definition for the zero mean normal cross correlation coefficient.

$$\rho = \frac{\sum_{x,y}(f(x,y)-\bar{f}_{m,n})(t(x-m,y-n)-\bar{t})}{\sqrt{\sum_{x,y}(f(x,y)-\bar{f}_{m,n})^2 \sum_{x,y}(t(x-m,y-n)-\bar{t})^2}} \quad (1)$$

In Equation (1) $\bar{f}_{m,n}$ denotes the mean value of *f(x, y)* within the area of the template t shifted to (m, n) which is calculated as

$$\bar{f}_{m,n} = \frac{1}{p \times q} \sum_{x=m}^{m+p-1} \sum_{y=n}^{n+q-1} f(x,y) \quad (2)$$

With similar notation $\bar{t}$ is the mean value of the template t. The denominator in equation (1) is the variance of the zero mean image function $f(x,y)-\bar{f}_{m,n}$ and the zero mean template function $t(x-m,y-n)-\bar{t}$. Due to this, the zero mean normalized cross correlation coefficient at (m,n) is independent to changes in brightness or contrast of the image which is related to mean value and standard deviation.

The desired position (m_{pos}, n_{pos}) of the pattern which is represented by t is equivalent to the position (m_{max}, n_{max}) at maximum value ρ_{max} of ρ(m, n). The zero mean normalized cross correlation method is more robust than other similarity measures like simple covariance or sum of absolute difference.

2.2 Hausdorff distance based image comparison

Hausdorff distance measures the extent to which each of model set lies near some point of an image set. The distance can be used to determine the degree of resemblance between two objects that are superimposed on one another. In this paper the second measure to improve the score of a match of template window to the search sub window.

A central problem in pattern recognition and computer vision is determining the extent to which one shape differs from another. Template matching can be viewed as a technique for determining the distance between shapes. In this algorithm Hausdorff distance is used as one of the measure to decide the relative position of a model and an image. The Hausdorff distance is a max-min distance as defined below.

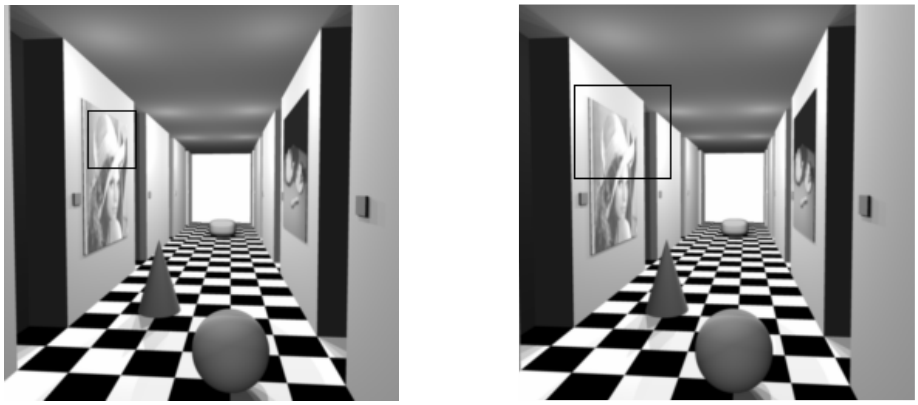


Figure 1 : Stereo image pair with left image with template window and right image with search window

Given two finite point sets A={ $a_1, a_2, \ldots\ldots, a_n$} and B= { $b_1, b_2, \ldots\ldots, b_n$}, the Hausdorff distance is defined as

$$H(A,B) = \max(h(A,B), h(B,A)) \quad (3)$$

Where

$$h(A,B) = \max_{a \in A} \min_{b \in B} \| a - b \| \quad (4)$$

And $\| . \|$ is L_2 or Euclidian norm. the function h(A,B) is called the direct Hausdorff distance from A to B. it identifies the point $a \in A$, that is farthest from any point of B and measures the distance from a to its nearest neighbour in B. Distance of each point of A to the nearest point of B is used for deciding ranking of each point of A. Thus the largest ranked point is considered as the distance (the most mismatched point of A). That means if h(A,B) = d then each point of A must be within distance d of some point of B and there is some point of A that is exactly distance d from nearest point of B (the most mismatched point). Hausdorff distance H(A,B) is the maximum of h(A,B) and h(B,A). It measures the degree of mismatch between two sets by measuring the distance of the point A that is farthest from any point of B. In this method of comparing shapes there is no explicit pairing of points of A with points of B. The computation time for H(A,B) will be o(mn) for two sets of size m and n.

2.3 Texture analysis using co-occurrence matrices

The co-occurrence has been described in image processing literature as gray-tone spatial dependence matrices (Haralick, Shanmugam and Dinstein, 1973). It has been widely used in texture analysis. The co-occurrence matrix in it core is a two-dimensional histogram of the occurrence of pairs of intensity values in a given spatial relationship.

Co-occurrence matrices are based on the relative frequencies $p(i, j)$ with which two pixels with a specified separation occur in the image, one with gray level i and another with gray level j. The separation is usually specified by distance vector s (d, θ). Pixel distance d and angular orientation θ are parameters of a particular co-occurrence matrices. Using different parameters, several matrices can be derived as shown in figure 2. The matrices obtained are then used for extraction of texture features. For the image of N gray levels, co-occurrence matrices cm can be obtained by estimating the pair wise statistics of pixel intensity. The size of cm is determined by the number of gray levels N in the input image.

The matrices cm are a functions of the angular relationship between the pixels as well as a function of the distance between them. An example of a 4x4 image with four gray levels and the computation of the co-occurrence matrices for d=1 and θ varying from 0° to 135° by 45° is given in figure 2.

In texture classification, individual elements of the co-occurrence are rarely used. Instead, features are derived from the matrix. A large number of textural features have been proposed starting with the original fourteen features described by Haralick, however only some of these are in wide use. The features which are used are listed as following:

Inverse Difference Moment

$$f_1 = \sum_{i,j} P(i,j) / 1 + (i-j)^2 \quad (5)$$

Contrast

$$f_2 = \sum_{i,j} \delta_{ij}^2 P(i,j) \quad (6)$$

0	0	1	1
0	0	1	1
0	2	2	2
2	2	2	3

(a) Image of 4 gray levels

	0	1	2	3 (Gray levels)
0	#(0,0)	#(0,1)	#(0,2)	#(0,3)
1	#(1,0)	#(1,1)	#(1,2)	#(1,3)
2	#(2,0)	#(2,1)	#(2,2)	#(2,3)
3	#(3,0)	#(3,1)	#(3,2)	#(3,3)

(Gray levels)

(b) General form of any co occurrence matrix #(i, j) is the number of times gray level i and j have been neighbors

$$cm(1,0^0) = \begin{bmatrix} 4 & 2 & 1 & 0 \\ 2 & 4 & 0 & 0 \\ 1 & 0 & 6 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix} \quad cm(1,45^0) = \begin{bmatrix} 4 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 \\ 0 & 2 & 4 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad cm(1,90^0) = \begin{bmatrix} 6 & 0 & 2 & 0 \\ 1 & 2 & 1 & 0 \\ 2 & 2 & 2 & 2 \\ 0 & 0 & 2 & 0 \end{bmatrix} \quad cm(1,135^0) = \begin{bmatrix} 2 & 1 & 3 & 0 \\ 1 & 2 & 1 & 0 \\ 3 & 1 & 0 & 2 \\ 0 & 0 & 2 & 0 \end{bmatrix}$$

Figure 2: Co-occurrence matrices for four given distance vectors (Haralick, Shanmugam and Dinstein, 1973)

Entropy

$$f_3 = -\sum_{i,j} P(i,j) \log P(i,j) \quad (7)$$

Correlation

$$f_4 = \sum_{i,j} (i-\mu_i)(j-\mu_j) P(i,j) / \sigma_i \sigma_j \quad (8)$$

Energy (angular second moment)

$$f_5 = \sum_{i,j} P(i,j)^2 \quad (9)$$

μ_i and μ_j are the means and σ_j and σ_j are the standard deviations of i and j respectively. These texture features are used for image matching by comparing the feature value of template window with the search window under consideration.

The five texture features are computed from the co-occurrence matrices of the template window and the search sub window and the difference between the feature coefficients is compared with a threshold which is empirically designed out for each feature. If the difference of the feature coefficients of three features are within the predefined range then the match is accepted. Sometimes cross-correlation technique fails to compare areas of smooth variance of intensities. Hence the accuracy of the match can be improved by texture analysis technique.

3. PROPOSED ALGORITHM

A composite method for area based image matching is proposed below:

1 Zero mean normal cross correlation coefficient between template window and search sub window is computed. The sub window having maximum cross correlation coefficient is considered as more perfect match for the template window under consideration. Best five matches are selected based on coefficient value and ranking of the matches is done based on cross correlation coefficient.

2 Further, edge map of template and search sub window is obtained by canny edge detector. Thus binary edge map of template and search sub window is derived.

3 The Hausdorff distance between the binary edge map of template and search sub window is used further for refining the best matches ; the lower the distance the best the match. The ranking of the best five matches of step one is done based on Hausdorff distance criteria.

4 Further texture analysis of the template and search sub window is done using co-occurrence matrix. The texture features are derived from the co-occurrence matrix using equations (5) to (9)

5 The steps 1 to 4 are performed for every basis images which are considered as separate template window.

Figure 3 shows the block diagram for the proposed algorithm

.

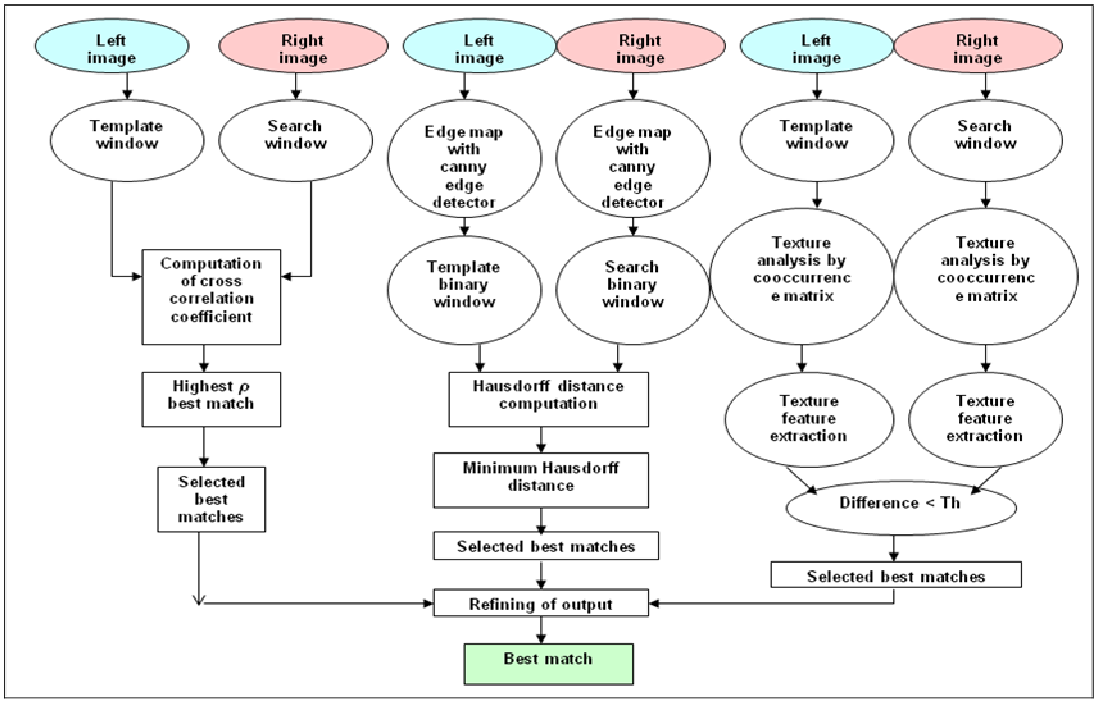


Figure 3 Block diagram for the proposed algorithm

4. RESULTS AND DISCUSSION

The proposed method is executed on the test stereo image pair shown in figure 1. As shown in the figure the left image shows the template window and the right image shows the larger search window. The size of the search window is decided by considering horizontal and vertical disparity measured by visual inspection of stereo image pair in figure 1.

As shown in figure 3, the block diagram for the proposed algorithm, the three methodologies used are zero mean normal cross correlation, Hausdorff distance and texture analysis. Parallel processing of the three methodologies can be incorporated for improving speed. Using zero mean normal cross correlation and considering largest five coefficient for the best matches for the given template window as shown in figure 4a. The matches found are shown in figure 4b. The minimum threshold of 0.8 is used for selecting the search sub window as a match. If the normalized cross correlation coefficient is less than 0.8 for all search sub window means that there is no match in the right image for the template window under consideration in the left image.

Further on the selected best matches and template window canny edge detector is applied and a set of binary images are obtained as shown in figure 4c. Computation of Hausdorff distance and ranking the matches based on Hausdorff distance is done. After the third step of the algorithm the best matches are filtered to the output as shown in figure 4d. The third refinement is done with texture analysis. Texture analysis is done on template window and selected search sub window of the figure 4d. The fourteen texture features are defined by (Haralick, R. M., Shanmugam, K. S., and Dinstein, I., 1973).

Figure 4a: Template window from left image from stereo image pair

Figure 4b: Selected sub widows as probable matches

Figure 4c: Binary edge images of sub windows of figre 4b

Figure 4d: Selected sub windows as matches for template window atfter applying Hausdorff distance criteria

Figure 4e: Best match selected after texture analysis

Out of the fourteen texture features the features entropy, contrast, inverse difference moment, correlation and energy (angular second moment) are computed and then best match is selected by comparing the feature values of template window and the refined output as shown in figure 4d. The best match is shown in figure 4e.

This procedure is repeated for every template window considered as a basis image. The first step is processed with a time complexity O(mn), where search window has pixels m X n. For further pass the comparison is done only on selected sub window making the execution faster.

The method removes most of the outliers and the accuracy of matching is improved compared to any single method giving a dense disparity map useful for 3D reconstruction. If parallel processing is done the method can be used in real time application like robotic vision. Figure 4e shows the best match.

5. CONCLUSION

Area based image matching algorithm offers a dense disparity map which is needed for 3-D construction of a scene from stereo images. In the proposed composite technique of image matching, three methodologies are used for refining the accuracy of the match. The normalized cross correlation coefficient is used for selecting best five matches and scores are given to them according to correlation coefficient value.

Further the scores of the matches are improved by Hausdorff distance technique. The third method that does texture analysis of template window and corresponding search sub window, improves the score of the matches. The best match is the one which is having highest score. This composite method of area based image matching improves the accuracy of the match and reduces number of outliers and gives dense disparity map. It can be used for real time application as the size of search window can be reduced by disparity range computed by visual inspection of stereo image pair reducing the computational cost. Further, refining of match with Hausdorff distance method and texture analysis method is done in linear time.

6. REFERENCES

Bernard, S.T. and Fischler. M.A. "Computational Stereo", *ACM Computing Surveys*, vol, 14, pp.553-572, 1982.

Canny, J., "A Computational Approach to Edge Detection",*IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 8, pp.679-714, 1986.

Egnal G., Wildes R., "Detecting binocular half-occlusions: empirical comparisons of four approaches", *Proc. Int. Conf. Comput. Vision Pattern Recognit.* (2) (2000) 466–473.

Fusiello A., Roberto V., and Trucco E., "Symmetric stereo with multiple windowing, *Int. Journal of Pattern Recognition and Artificial Intelligence,* 14:1053-1066, 2000.

Fusiello A., Trucco E., Verri E., A compact algorithm for rectification of stereo pairs, Machine Vision Appl. 12 (1) (2000) 16–22.

Haralick, R.M. and Shapiro L.G., “Computer and Robotic Vision”, in *MA: Addition-Wesley,* Volume: 2, Chapter: 16, 1993

Haralick, R. M., Shanmugam, K. S., and Dinstein, I., “Textural Features for Image Classification”, *IEEE International Conference on Systems, Man, and Cybernetics,* Vol. SMC-3, No.6, pp 610-621, 1973

http://www-cgrl.cs.mcgill.ca/~godfried/teaching/cg-projects/98/normand/main.html

Huttenlocher D. P., Kedam K “Efficiently computing the Hausdorff distance for point sets under translation,”, *Proc. Sixth ACM Symp. Computat. Geometry,* 1990, pp.340-349.

Lucas B. D., Kanade T “An Iterative Technique with an application to Stereo Vision”, *IJCAI*. 1981.

Muhlmann K., Maier D., Hesser J., Manner R., “Calculating dense disparity maps from color stereo images, an efficient implementation”„ *Int. J. Comput. Vision* 47 (1–3) (2002) 79–88.

Peleg A Weiser U., MMX technology extension to the intel architecture, *IEEE Micro* 16 (4) (1996) 42–50.

Pratt, W.K , "Correlation Techniques of Image Registration.", *IEEE Transactions on Aerospace and Electronic Systems,* vol.10, no. 3, pp. 353-358, 1974.

Sun, C., "A Fast Stereo Matching Method", *Digital Image Computing: Techniques and Application,* pp. 95-100, Massey University, Auckland, New Zealand, 1997.

Trucco E.,Verri A., “Introductory techniques for 3D computer vision”, *Prentice Hall, Englewood cliffs, NJ*, 1998.

Zitova, B. and Flusser J, “Image Registration Methods: A Survey”, *Image and Vision Computing,* Vol 21, No. 11, pp 977-1000, 2003.

AN EXPERIMENTAL STUDY ON REGISTRATION THREE-DIMENSIONAL RANGE IMAGES USING RANGE AND INTENSITY DATA

Cihan ALTUNTAS

Selcuk University, Engineering and Architectural Faculty, Geomatic Engineering, 42075, Selcuklu, Konya, Turkey
caltuntas@selcuk.edu.tr

KEY WORDS: Point cloud, Registration, Intensity image, Range image, Automation, Keypoint.

ABSTRACT:

Laser scanner is noncontact instrument to measurement of spatial data. It measures object surfaces as point series and visualize as point cloud. One of the important steps on processes of laser scanning data is the registration of point clouds relation to common coordinate system. Many interactive and automatic methods have been developed for point cloud registration so far. The automatic methods are applied with range data of laser scanner or image data of sensor combination camera. The registration by range data is mostly depend object geometry. If scan surface is deprived from geometrical details, conjugate points can not be found to compute registration parameters between point clouds. In that case, intensity data of laser points can be used for registration. In this study, intensity image was created from laser scanner data and the registration parameters were computed with keypoints extracted by SIFT method from these images. The results were also compared with the iterative closest point (ICP) method.

1 INTRODUCTION

Today many applications require three-dimensional (3D) modelling of object or scene. The latest method for 3D modelling is laser scanning and it has been widely used for documentation of cultural heritage, deformation measurement, reverse engineering, virtual reality, architectural planning and scene monitoring. Laser scanner is non contact instrument to measurement spatial data. In addition, intensity data backscattered laser beam from the scan point is recorded. Furthermore, color data (RGB) can be recoded via the image of integrated camera.

Many scanning are performed from different station to obtain occlusion free 3D object model. Laser scanner data (point cloud) is in local coordinate system centre of which is the laser scanner. In this case, all point clouds must be registered into common coordinate system to visualizing 3D model of the object. Generally first point cloud was selected as a reference and the others are registered into its coordinate system. Many methods have been developed for registration of point clouds. The most popular method is iterative closest point (ICP) method (Chen and Medioni, 1992; Besl and McKay, 1992). Another one is least square 3D image matching method (Guen and Akca, 2005). In addition, registration can be performed with object details extracted from point clouds (Deveau et al., 2004; Briese and Pfeifer, 2008). The registration methods of point clouds were investigated with details in Salvi et al. (2007).

The methods mentioned above needs initial registration parameters to perform. Thus automatic registration of point clouds has still been research area. The automatic methods are performed with coarse-to-fine strategy. Initially, coarse registration parameters have been computed by different methods and than fine registration is performed by ICP. Also, automatic methods were executed with range image of laser scanner data in literature. Range image and real camera image were evaluated by photogrammetric in Aquilera et al. (2009). Camera position was estimated relation to the laser scanner and image texture data was mapped with point cloud. In another study, range images were created from point clouds and keypoints extracted by SIFT operator. The registration parameters were computed by laser coordinates of key points (Barnea and Filin, 2008; Körtgen 2007; Sharp, 2002). Bendels et al. (2004) was performed automatic multi-view registration with range and intensity image. Keypoints were extracted by SIFT (Lowe, 2003) method and laser coordinates of keypoints were computed. Then points were matched by RANSAC (Fischler and Bolles, 1981) and registration parameters were computed. In another study, range image was created from point cloud and object planar surface was extracted from point cloud. The registration parameters were computed by planar surfaces and range image (Dold and Brenner, 2004). But, object details can not be selected from them since the range image was created from spatial range data. Range image dose not include intensity backscattered laser beam from the scan points. The intensity data is represent brightening and color details of the surface. In this study, point clouds were registered by intensity image which was created from laser scanner data.

2. THE REGISTRATION METHOD

Laser scans have been made as overlapping beginning of the first point cloud. In that case, to combine all point clods, the registration must be performed relation to reference point clouds. The registration parameters have been computed by laser scanning points with different techniques in overlapping area. In this study, pair-wise registration was performed by intensity image created from laser scanner data. Keypoints from intensity images were extracted and matched by SIFT method. After outlier points were detected by statistical method, point cloud coordinates for each conjugate keypoints were determined using range image data. The registration parameters were computed with these corresponding points and residuals on SIFT points were computed. Then the results of the method were compared with the ICP.

2.1 Itensity and Range Image

Intensity and range image has two-dimensional (2D) which has the same view with laser scanner. In this image, 3D laser data is represented by two-dimensional image pixels. At first, image

marges are defined by min and max angles of point cloud data. The image width is limited by horizontal angle as min angle is right and max angle is left side. As a similar, the image height is limited by vertical angle as min angle is bottom and max angle is top side. Then the grid which is present the pixel of the intensity image is created within image limit. An each excel of the grid is pixel of the image and pixel size is described by the angle correspond to laser scanning point steps of the station. Pixel size angle is a function of mean scan distance and point steps (Eq.1). Nevertheless, because of the variable scan distance, pixel size angle must be select small from computed value. The pixels are colored by intensity and range data for intensity and range image respectively. In addition, pixel size angles and range data are also recorded into the file. Each laser scanning point is represented by one pixel on intensity image. If the pixel is not corresponded the laser point, its value is interpolated from near pixels. Grey tones of the range image are computed by range data. If the image is created as 8 bite, grey tones for min and max ranges will be lies between 0-255.

Pixel size angle=$\tan^{-1}$(laser point space/mean scan distance) (1)

2.2 Keypoint Detection

Many methods such as Haris, FAST, SUSAN, Forstner and SIFT have been used for keypoints detection (Jzayeri and Fraser, 2010). SIFT is extensively used for keypoint detection. Because it can detect keypoints even if scale, rotation and color are different between the images. For image matching, SIFT features are first extracted from reference image and stored in database. The new image is matched by comparing each feature from the new image to this previous database. The candidate matching features is found by Euclidean distance of their feature vector. In this study two intensity images have different scale, rotation and angle. However extracted keypoints include outlier points. This outlier points must be removed before estimation of the registration parameters.

Figure 1. Laser scanning

2.3 Outlier Detection

Point cloud coordinates of keypoints are computed by horizontal and vertical angle and range data recorded into the file of range image. After laser coordinates of keypoints were computed, the registration parameters between point clouds are computed by least four point selected by RANSAC (Fischler and Bolles, 1981) method. Then the second point cloud was registered into the first point cloud with these parameters. Afterwards, residuals on keypoints are computed and the points which have the biggest residuals are eliminated. This computation is iterated until the biggest residual of coordinates will be small from inherent accuracy (8mm@100m for ILRIS 3D) of laser scanner (Kang, 2008).

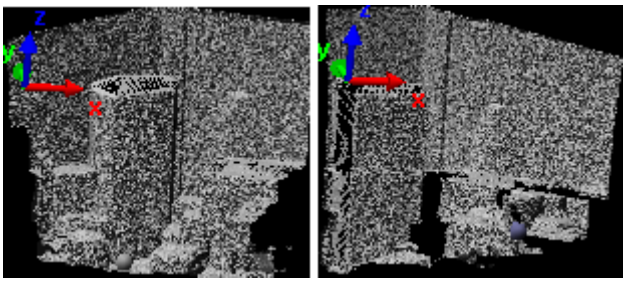


Figure 2. Overlapping point clouds from two stations.

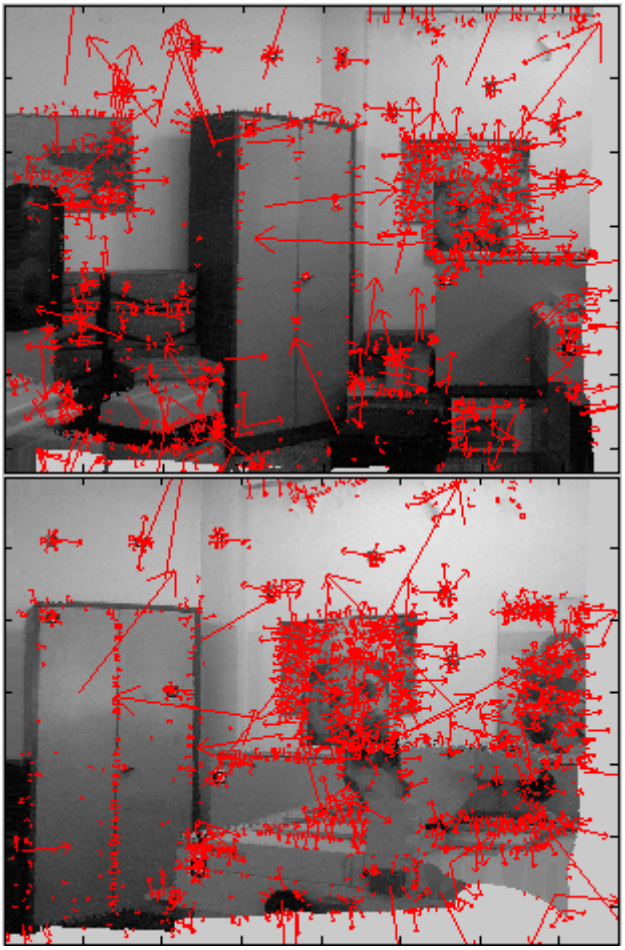


Figure 3. SIFT Keypoints with scale and rotations. 2121 and 3152 keypoints were extracted from intensity images of the first and second point clouds respectively.

3. EXPERIMENT

Laser scannings were performed by ILRIS 3D laser scanner from two stations as overlapping in our photogrammetry laboratory (Figure 1). The range of laser scanning is about 6.5m and 5 m respectively. The surfaces were scanned 3mm point spaces and 411000 and 233000 points were collected from two

tations respectively (Figure 2). Two-dimensional intensity and range image was created for each scan. The pixel size angle which was correspond laser point spaces were computed 0.03 grad. The first intensity image has 1536x1260 pixels and the second has 1548x1300 pixels. The SIFT method was executed by Matlab code developed Lowe (2005). Respectively 2121 and 3152 keypoints were extracted from the intensity images (Figure 3) and 78 keypoints were matched (Figure 4).

After the compute laser coordinates of keypoints, the registration parameters between point clouds were computed with four points which were selected by RANSAC in Matlab program. The second point cloud was registered into the first point cloud with these registration parameters. Then residuals on keypoints were computed and keypoints which have the biggest residuals are eliminated. At last, 30 keypoints were remained which have small residual than 8mm (inherent accuracy is 8mm@100m). In addition, the registration was executed by ICP method and residuals are computed on 30 keypoints (Table 1).

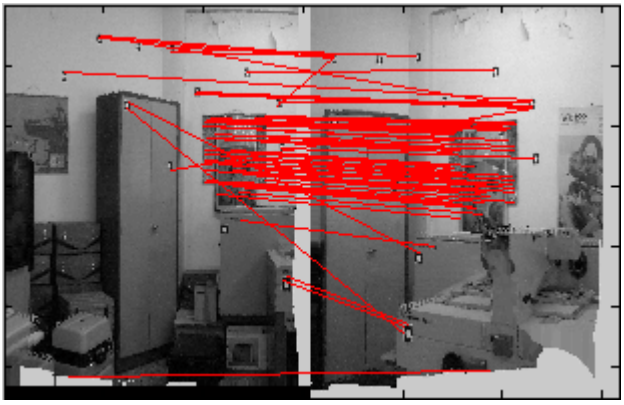


Figure 4. Total 78 keypoints were matched by SIFT parameters.

Tablo 1: After the registration by ICP and intensity image methods, min and max residuals on keypoint coordinates

Method		dx (cm)	dy (cm)	dz (cm)	ds=sqrt(dx^2+dy^2+dz^2) (cm)
ICP	Min	-0.15	0.04	-0.04	0.16
	Max	-0.78	1.25	-0.62	1.60
Intensity Image	Min	0.12	-1.04	0.41	1.12
	Max	0.94	-2.72	0.50	2.92

4. DISCUSSION

SIFT method is more appropriate method for keypoint extraction from like these images. Outlier detection can be performed different methods. In this study, threshold value for outlier detection may be change according to the laser instrument and expected precise. The results of the registration by intensity image are near the ICP method (Table 1). This method can be used for automatic registration of point clouds as independent object geometry. On the other hand, the method can be used to compute initial registration parameters for automatic registration. Matched keypoints are include a little false matched points. Therefore the registration parameters to perform ICP can be computed by keypoints without outlier detection.

5. CONCLUSION

In this study, pair-wise registration was performed by using intensity image. Keypoints were extracted by SIFT method and the registration parameters were computed laser coordinates of them. The registration is also performed by ICP and the coordinate residuals on keypoints were compared. As a consequence, the registration by intensity image gave near results with the ICP. Moreover, the method can be used for pair-wise automatic registration of point clouds.

ACKNOWLEDGEMENT

The author thanks to Prof. Norbert PFEIFER for valuable comments during study on Vienna University of Technology.

REFERENCES

Aquilera, D.G., Gonzalvez, P.R. and Lahoz, J.G., 2009. An automatic procedure for co-registration of terrestrial laser scanners and digital cameras, *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(2009), pp. 308-316.

Barnea, S., Filin, S., 2008. Keypoint based autonomous registration of terrestrial laser scanner point-clouds, *ISPRS Journal of Photogrammetry and Remote Sensing*, 63 (2008), pp. 19-35.

Bendels, G.H., Degener, P., Wahl, R., Körtgen, M., Klein, R., 2004. Image-based registration of 3D-range data using feature surface elements, *The 5th International Symposium on Virtual Reality, Archeology and Cultural Heritage VAST (2004)*, Oudenaarde, Belgium, pp. 115-124.

Besl, P.J., McKay, N.D., 1992. A method for registration of 3-D shapes, *IEEE Transsactions on Pattern Analysis and Machine Intelligence*, 14(2), pp. 239-256.

Briese, C., Pfeifer, N., 2008. Line based reconstruction from terrestrial laser scanning data, *Journal of Applied Geodesy*, 2(2008), pp. 85-95.

Chen, Y., Medioni, G., 1992. Object modelling by registration of multiple range images, *Image and Vision Computing*, 10(3), pp. 145–155.

Deveau, M., Deseilligny, M.P., Paparoditis, N., Chen, X., 2004. Relative laser scanner and image pose estimation from points and segments, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (IAPRS)*, 35(B/3), Istanbul, pp. 1130-1135.

Dold, C., Brenner, C., 2004. Automatic matching of terrestrial laser scan data as a basis fort he generation of detailed 3D city models. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (IAPRS), 35(B/3), Istanbul, 12-23 July, 2004, pp. 1091-1096.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, *Communications Association and Computing Machine*, 24(6), pp. 381-395.

Gruen, A., Akca, D., 2005. Least squares 3D surface and curve matching, *ISPRS Journal of Photogrammetry and Remote Sensing*, 59(2005), pp. 151-174.

Kang, Z., 2008. Automatic registration of terrestrial point cloud using panoramic reflectance images, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (IAPRS), 37(B/5), Beijing, pp. 431-35.

Körtgen, M., 2007. Robust automatic registration of range images with reflectance, *Diplomarbeit, Vorgelegt am Institut für Informatic*, Rheinische Friedrich-Wilhelms Universitat Bonn, 89 pages.

Lowe, D., 2003. Distinctive image features from scale-invariant keypoints. *International journal of Computer Vision*, Vol.20, pp. 91-110.

Jzayeri, I., Fraser, C., 2010. Interest operators for feature-based matching in close range photogrammetry, *The Photogrammetric Record*, 25(129), pp. 24-41.

Salvi, J., Matabosch, C., Fofi, D., Forest, J., 2007. A review of recent range image registration methods with accuracy evaluation, *Image and Vision Computing*, 25 (2007), pp. 578-596.

Sharp, G.C., Lee, S.W., Wehe, D.K., 2002. ICP registration using invariant features, *IEEE Trans Saction on Pattern Analysis and Machine Intelligence*, 24(1), pp. 90-102.

Lowe, D., 2005. Demo Software: SIFT Keypoint Detector. http://www.cs.ubc.ca/~lowe/keypoints/ (Accessed 20 May, 2011).

SEMI-AUTOMATIC IMAGE-BASED CO-REGISTRATION OF RANGE IMAGING DATA WITH DIFFERENT CHARACTERISTICS

Ma. Weinmann, S. Wursthorn, B. Jutzi

Institute of Photogrammetry and Remote Sensing, Universität Karlsruhe
Kaiserstr. 12, 76128 Karlsruhe, Germany
{martin.weinmann, sven.wursthorn, boris.jutzi}@kit.edu

KEY WORDS: Accuracy, automatic, evaluation, close-range, range imaging, active sensing, structured light.

ABSTRACT:
Currently, enhanced types of active range imaging devices are available for capturing dynamic scenes. By using intensity and range images, data derived from different or the same range imaging devices can be fused. In this paper, an automatic image-based co-registration methodology is presented which uses a RANSAC-based scheme for the Efficient Perspective-n-Point (EPnP) algorithm. For evaluating the methodology, two different types of range imaging devices have been investigated, namely Microsoft Kinect and PMD [vision] CamCube 2.0. The data sets captured with the test devices have been compared to a reference device with respect to the absolute and relative accuracy. As the presented methodology can cope with different configurations concerning measurement principle, point density and range accuracy, it shows a high potential for automated data fusion for range imaging devices.

1. INTRODUCTION

The capturing of 3D information about the local environment is still an important topic as this is a crucial step for a detailed description or recognition of objects within the scene. Most of the current approaches are based on the use of image or range data. By using passive imaging sensors like cameras, the respective 3D information is obtained indirectly via textured images and stereo- or multiple-view analysis with a high computational effort. These procedures are widely used, but they depend on scenes with adequate illumination conditions and opaque objects with textured surface. Besides, the distances between sensor and object, between the different viewpoints of an imaging sensor and between the sensors of the stereo rig, in the case of using a stereo camera, should be sufficiently large in order to obtain reliable 3D information.

In contrast to the photogrammetric methods, terrestrial laser scanner (TLS) devices allow for a direct and illumination-independent measurement of 3D object surfaces (Shan & Toth, 2008; Vosselman & Maas, 2010). These scanning sensors capture a sequence of single range values on a regular spherical scan grid and thus accomplish a time-dependent spatial scanning of the local environment. Hence, the scene contents as well as the sensor platform should be static in order to reach an accurate data acquisition.

For an adequate capturing of dynamic scenes given for instance by moving objects or a moving sensor platform, it is essential to obtain all and dense 3D information about the local environment at the same time. Recent developments show that enhanced types of active imaging sensors have started to meet these requirements. Suitable for close-range perception, these sensors allow for simultaneously capturing a range image and a co-registered intensity image while still maintaining high update rates (up to 100 releases per second). However, the non-ambiguity range of these sensors is only several meters and depends on the modulation frequency. This problem can currently only be tackled by using active imaging sensors based on different modulation frequencies (Jutzi, 2009; Jutzi, 2011). Besides, the measured intensity strongly depends on the wavelength (typically close infrared) of the laser source as well as on the surface characteristic. Various studies on range imaging focus on hardware and software developments (Lange, 2000), geometric calibration (Reulke, 2006; Kahlmann et al., 2007; Lichti, 2008) and radiometric calibration (Lichti, 2008).

Nowadays, many approaches for capturing single 3D objects are still based on the use of coded structured light. In Salvi et al. (2004), different strategies for pattern codification are summarized and compared. In general, all these strategies are based on the idea of projecting a coded light pattern on the object surface and viewing the illuminated scene. Such coded patterns allow for a simple detection of correspondences between image points and points of the projected pattern. These correspondences are required to triangulate the decoded points and thus obtain the respective 3D information. For real-time applications or dynamic scene acquisition, it is essential to avoid time-multiplexing methods as these usually depend on the successive projection of different binary codes. Very simple patterns with inexpensive hardware requirements which are also suitable for dynamic scenes can for example be established via dot patterns. Using regular dot patterns for measuring surfaces of close-range objects by considering the images of several CCD cameras has been presented in Maas (1992) and offers advantages like redundancy, reliability and accuracy without the need of a priori information or human interaction. The idea of using dot patterns has further been improved and currently, new types of sensors (e.g. the Microsoft Kinect device developed by PrimeSense) use random dot patterns of projected infrared points for getting reliable and dense close-range measurements in real-time.

Using the new types of active imaging sensors is well-suited for dynamic close-range 3D applications, e.g. like the autonomous navigation of robots, driver assistance, traffic monitoring or tracking of pedestrians for building surveillance. Therefore, it is important to further investigate the potentials arising form these sensor types.

In this paper, a method for semi-automatic image-based co-registration of point cloud data is proposed, as an accurate range measurement with a reference target for a large field-of-view is technically demanding and can be expensive. For an automatic image-based algorithm, various general problems have to be tackled, e.g. co-registration, camera calibration, image transformation to a common coordinate frame and resampling. With the range imaging devices (e.g. PMD [vision] CamCube 2.0 and Microsoft Kinect) test data is captured and compared to reference data derived by a reference device (Leica HDS6000). The general framework focuses on an image-based co-registration of the different data types, where keypoints are detected within each data set and the respective transformation

parameters are estimated with a RANSAC-based approach to camera pose estimation using the Efficient Perspective-n-Point (EPnP) algorithm. Additionally, the proposed algorithm can as well be used to co-register data derived from different or the same ranging devices without adaptations. This allows for fusing range data in form of point clouds with different densities and accuracy. A typical application could be the completion and densification of sparse data with additional data in a common coordinate system. After this co-registration, the absolute and relative range accuracy of the range imaging devices are evaluated by experiments. For this purpose, the data sets captured with the test devices over a whole sequence of frames are considered and compared to the data set of a reference device (Leica HDS6000) transformed to the local coordinate frame of the test device. The results are shown and discussed for an indoor scene.

The remainder of this paper is organized as follows. In Section 2, the proposed approach for an image-based co-registration of point clouds and a final comparison of the measured data is described. The configuration of the sensors and the scene is outlined in Section 3. Subsequently, the captured data is examined in Section 4. The performance of the presented approach is tested in Section 5. Then, the derived results are evaluated and discussed in Section 6 and finally, in Section 7, the content of the entire paper is concluded and an outlook is given.

2. METHODOLOGY

For comparing the data captured with a range imaging device to the data captured with a laser scanner which serves as reference, the respective data must be transformed into a common coordinate frame. Therefore, the change in orientation and position, i.e. the rotation and translation parameters between the different sensors, has to be estimated. As illustrated in Figure 1, it is worth analyzing the data after the data acquisition.

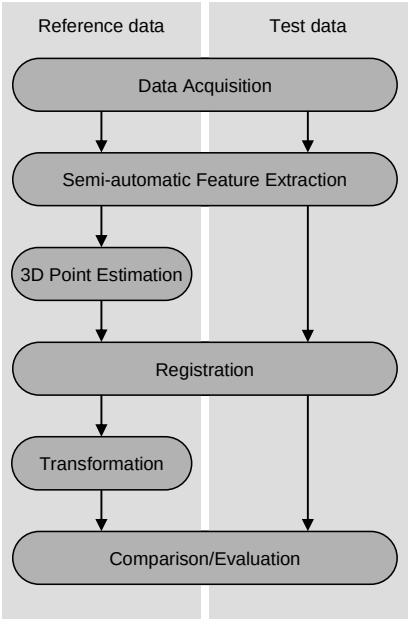


Figure 1. Processing chain of the proposed approach.

The laser scanner provides data with high density and high accuracy in the full range of the considered indoor scene, whereas the range imaging devices are especially suited for close-range applications. Hence, the rotation and translation parameters can be estimated via 3D-to-2D correspondences between 3D points derived from the TLS measurements and 2D image points of the respective range imaging sensor. These 3D-to-2D correspondences are derived via a semi-automatic selection of point correspondences between the intensity images of the laser scanner and the test device, and built by combining the 2D points of the test device with the respective interpolated 3D information of the laser scanner. In Moreno-Noguer et al. (2007) and Lepetit et al. (2009), the Efficient Perspective-n-Point (EPnP) algorithm has been presented as a non-iterative solution for estimating the transformation parameters based on such 3D-to-2D correspondences. As the EPnP algorithm takes all the 3D-to-2D correspondences into consideration without checking their reliability, it has furthermore been proposed to increase the quality of the registration results by introducing the RANSAC algorithm (Fischler & Bolles, 1981) for eliminating outliers and thus reaching a more robust pose estimation. Using the estimated transformation parameters, the reference data is transformed into the local coordinate frame of the test device. This part of the proposed approach is comparable to the coarse registration presented in Weinmann et al. (2011). Finally, the estimated transformation allows for comparing the captured data.

3. CONFIGURATION

To validate the proposed methodology, a configuration concerning sensors and scene has to be utilized.

3.1 Sensors

For the experiments, two different range imaging devices were used as test devices and a terrestrial laser scanner as reference device.

3.1.1 Range imaging device - PMD [vision] CamCube 2.0

With a PMD [Vision] CamCube 2.0, various types of data can be captured, namely the range and the intensity, where the intensity can be distinguished in active and passive intensity. The measured active intensity depends on the illumination emitted by the sensor and the passive intensity on the background illumination (e.g. sun or other light sources). The data can be depicted as image with an image size of 204 x 204 pixels. A field-of-view of 40° x 40° is specified in the manual.

Currently, the non-ambiguity which is sometimes called unique range is less than 10 m and depends on the tunable modulation frequency. This range measurement restriction can be improved by image- or hardware-based unwrapping procedures in order to operate as well in far range (Jutzi, 2009; Jutzi, 2011).

For the experiments the hardware-based unwrapping procedures were utilized, where modulation frequencies of 18 MHz and 21 MHz were selected for maximum frequency discrimination. The integration time was pushed to the maximum of 40 ms in order to gain a high signal-to-noise ratio for the measurement. In this case, saturation could appear in close range or arise from object surfaces with high reflectivity. All measurement values were captured in raw mode.

3.1.2 Range imaging device - Microsoft Kinect

The Microsoft Kinect device is a game console add-on which captures disparity and RGB images with a frame rate of 30 Hz. Originally, the disparity images are used to track full body skeleton poses of several players in order to control the game

play. The device has a RGB camera, an IR camera and a laser-based IR projector which projects a known structured light pattern of random points onto the scene. IR camera and IR projector form a stereo pair. The pattern matching in the IR image is done directly on-board resulting in a raw disparity image which is read out with 11 bit depth. Both RGB and disparity image have image sizes of 640 x 480 pixels. The disparity image has a constant band of 8 pixels width at the right side which supports speculation (Konolige & Mihelich, 2010) of a correlation window width of 9 pixels used in the hardware-based matching process. For the data examination, this band has been ignored, which yields a final disparity image size of 632 x 480 pixels.

Camera intrinsics, baseline and depth offset have been calibrated in order to transform the disparities to depth values and to register RGB image and depth image. The horizontal field-of-view of the RGB camera is with 63.2° wider than the field-of-view of the IR camera with 56.2°. Considering the stereo baseline of 7.96 cm, known from calibration, the range is limited. The Kinect device is based on a reference design (1.08) from PrimeSense, the company that developed the system and licensed it to Microsoft. In the technical specifications of the reference design, an operation range for indoor applications from 0.8 to 3.5 m is given.

Figure 2. Range imaging devices: Microsoft Kinect (left) and PMD[vision] CamCube 2.0 (right).

3.1.3 Reference device - Leica HDS6000

The Leica HDS6000 is a standard phase-based terrestrial laser scanner with survey-grade accuracy (within mm range) and a field-of-view of 360° x 155°, and the full captured image size is 2500 x 1076 pixels. Hence, the angular resolution is approximately 0.14°.

3.2 Scene

A data set of a static indoor scene was recorded with the stationary placed sensors mentioned above. In Figure 3, a photo of the observed scene is depicted. For the environment no reference data concerning the radiometry or geometry was available. Hence, the scene is more suited for investigating the quality of the test devices at different levels of distance, even beyond the sensor specifications, where it will be seen that the captured information might eventually still be suitable.

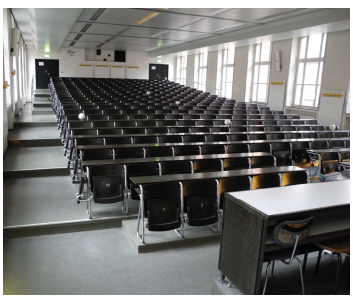


Figure 3. RGB image of the observed indoor scene.

4. DATA EXAMINATION

In this section, the semi-automatic feature extraction by an operator, the transformation of the data into a common coordinate system and finally, the resampling of the data into a proper grid is described.

4.1 Semi-automatic feature extraction

For an efficient registration process, it has proved to be suitable to establish pairs of points, each consisting of a 3D point representing information derived from the reference data and a 2D point representing the image coordinates measured in the image information of the test device (Weinmann et al., 2011). Based on these 3D-to-2D correspondences, the co-registration can be carried out via the Efficient Perspective-n-Point (EPnP) algorithm which has recently been presented as a fast and accurate approach to pose estimation.

Hence, the image coordinates of the control points have been measured manually and with sub-pixel accuracy in the passive intensity image of the test devices, which has been undistorted and mapped to the depth image, as well as in the image of the reference device. Subsequently, the corresponding 3D object coordinates have been determined based on the reference data by interpolation as the measured 3D information is only available on a regular grid.

The proposed approach consisting of EPnP and RANSAC has been used to estimate the exterior orientation of both test devices in relation to the reference data. Table 1 shows the resulting reprojection errors, the number of all determined control points and the number of control points selected by the RANSAC algorithm. The low percentage of utilized control points is only slightly influenced by a low quality of the manual 2D measurement but rather by the range information itself. As distinctive 2D control points are selected first which are located at corners or blobs, the respective interpolated 3D information may abruptly change and thus not always be reliable.

	Reprojection error [pixel]	Number of control points	
		available	used
PMD [vision] CamCube 2.0	0.693	13	7
Microsoft Kinect	0.328	21	11

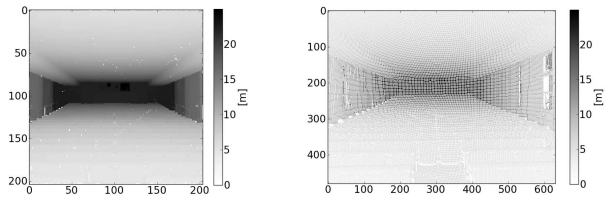
Table 1. Quantity and quality of the utilized control points.

4.2 Converting range to depth images

Once the transformation parameters between reference and test device are estimated, it is possible to check how 3D points measured with the reference device are projected onto the image plane of a virtual camera with the same intrinsic parameters as the test device. Using homogeneous coordinates, this transformation can be formulated as

$$\mathbf{x}'_{Ref} = \mathbf{K}\left[\mathbf{R} \mid \mathbf{t}\right]\mathbf{X}'_{Ref} \qquad (1)$$

where **K** is the calibration matrix of the virtual camera, **R** the estimated rotation matrix and **t** the estimated translation vector. If a pixel in this virtual camera image is assigned more than one of the 3D points, the mean values of the respective points are used. Resulting from this, resampled synthetic depth images can be created, which are shown in Figure 4 for using the same calibration matrices as those of the two test devices.



a b

Figure 4. Synthetic depth images: a) PMD [vision] CamCube 2.0, b) Microsoft Kinect.

Due to the given lower angular resolution of the reference device (0.14°) in comparison to the test device Microsoft Kinect (0.09°), artifacts from resampling can be observed in the synthetic depth image in Figure 4b. The test device PMD [vision] CamCube 2.0 records the data with an angular resolution of 0.20° which is lower than the angular resolution of the reference device. For that reason, the synthetic depth image in Figure 4a is without artifacts. Thus, the depth values of the different devices can easily be compared to the depth values of the reference device.

The absolute accuracy is given by the *depth difference*

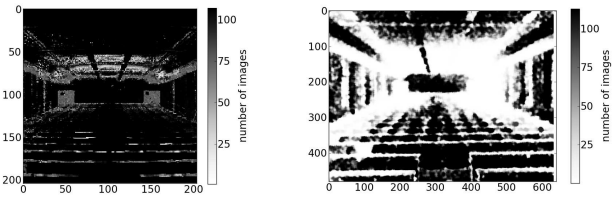
$$\Delta_z(x',y') = z_{Ref}(x',y') - \overline{z}(x',y') \,, \qquad (2)$$

which is calculated by the difference between reference depth z_{Ref} derived from the reference device and the mean value $\overline{z}$ derived from at least 100 single measurements captured by the investigated range imaging device over a time sequence.

Then, the relative accuracy is given by the *standard deviation of the depth difference* σ_z.

5. ANALYSIS RESULTS

First over 100 images of the static scene have been captured with both fixed devices, and these images are represented by a stack of images. Unreliable measurement values, resulting from noise effects, yield less than 100 values and have been masked out. The remaining reliable measurement values are utilized for further analysis. The number of reliable measurements depicted by gray values is shown in Figure 5.



a b

Figure 5. Number of available measurement values: a) PMD [vision] CamCube 2.0, b) Microsoft Kinect.

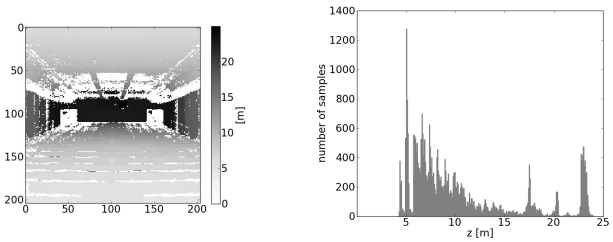
For the range imaging device PMD [vision] CamCube 2.0, a total number of 33835 reliable pixels (81%) meets our constraints. For the range imaging device Microsoft Kinect, the maximum raw disparity of 2047 (at 11 bits) has been masked out additionally, which yields a total number of 104478 reliable pixels (34%).

From the reliable values, the mean and the standard deviation of the depth have been calculated.

5.1 Range imaging device - PMD [vision] CamCube 2.0

In Figure 6a, the mean depth obtained with the PMD [vision] CamCube 2.0 is visualized. Unreliable measurement values, which are represented with white color, appear at the polished surfaces in the foreground mainly on the left side where the incidence angle to the surface is steep, resulting in uncertainties (Figure 7a). Further unreliable measurement values can be observed on the dark colored and polished doors in the back of the room. These outliers occur due to the low reflectivity or specular surface characteristic which can result in multipath measurements.

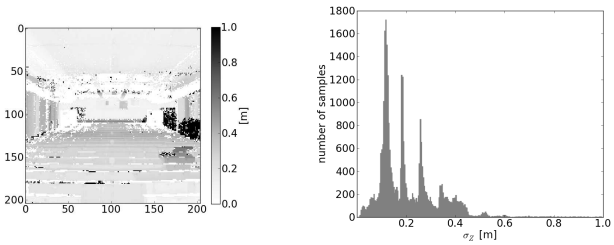
The depth values are spread over an interval from 4.16 to 24.94 m. Figure 6b shows a histogram of the estimated mean depth. Due to a maximum distance to the central wall at the back of the room of about 23 m, absolute range values above this distance are erroneous.



a b

Figure 6. Mean depth: a) gray-coded image, b) histogram.

To the mean depth mentioned above, the corresponding standard deviation is shown in Figure 7, where most of the values are below σ_z with 0.5 m. The standard deviation increases slightly with depth and a maximum of 4.62 m can be observed in the data.



a b

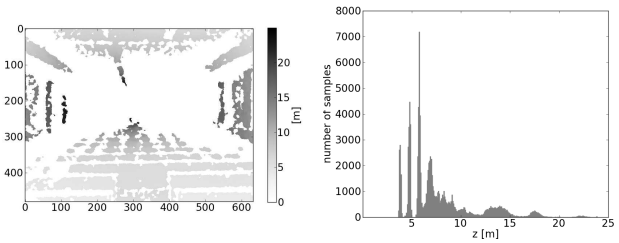
Figure 7. Standard deviation of the depth: a) gray-coded image, b) histogram.

5.2 Range imaging device - Microsoft Kinect

In Figure 8a, the mean depth obtained with the Microsoft Kinect is visualized. Obviously, the operation range has been exceeded in the selected scene. Hence, the wall at the back of the room is completely missing, because the maximum raw disparity values have been filtered out (compare Figure 5b to 8a). However the remaining depth measurements still show varying distances to different rows of chairs indicating the rough structure of the scene. The depth values are within an interval from 3.61 to 23.86 m (Figure 8b). This statement supports a use of this test device for densifying sparse depth measurements far beyond the sensor specification.

The object size with its surface direction, where the pattern is projected on, and the correlation window size lead to limitations

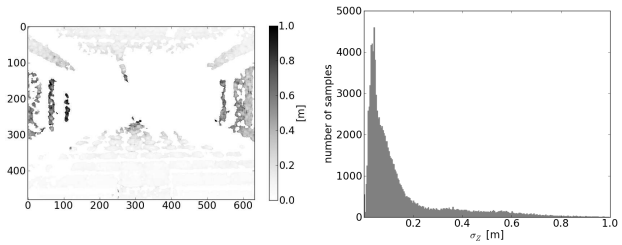
with respect to the spatial resolution of the depth image. For instance, there is no clear partition in depth for the more distant rows of chairs compared to the PMD [vision] CamCube 2.0, where depth stepping of rows can be resolved up to the last row.



a b

Figure 8. Mean depth: a) gray-coded image, b) histogram.

As can be seen in Figure 9a, the standard deviation increases with depth, where most of the values are below σ_z with 0.2 m and a maximum of 1.41 m is given.

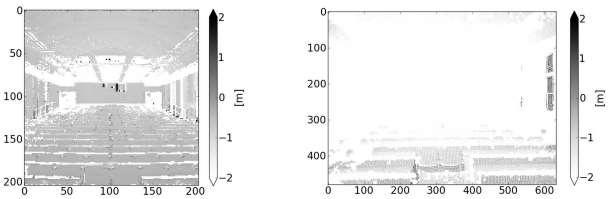


a b

Figure 9. Standard deviation of the depth: a) gray-coded image, b) histogram.

6. EVALUATION AND DISCUSSION

Finally, the derived depth differences are evaluated and discussed by calculating the mean depth and the standard deviation of the depth. In Figure 10, the depth differences per pixel are shown and in Figure 11, the corresponding density distributions are depicted.

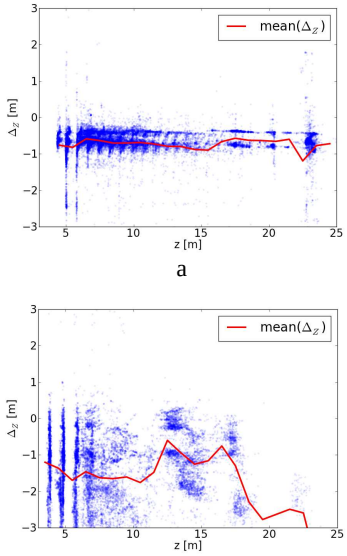


a b

Figure 10. Depth difference between the data of reference and test device: a) PMD [vision] CamCube 2.0, b) Microsoft Kinect.

Homogenous areas can be stated for the PMD [vision] CamCube 2.0 in Figure 10a. These areas represent a systematic range shift, where the range measurement tends to be too close to the sensor.

Concerning the reliable pixels over the scene depth, 25109 depth difference values (74%) are within the interval [-1,0] m. The standard deviation of the depth difference might depend on the signal-to-noise ratio of the measurement. Due to the inverse square law concerning range dependency of the received light power, the estimated mean value in Figure 12a follows this trend.

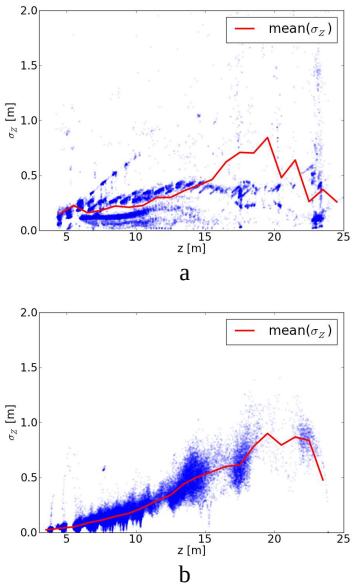


a

b

Figure 11. Density distribution (blue dotted) and mean (red solid) of depth z versus depth difference Δ_z: a) PMD [vision] CamCube 2.0, b) Microsoft Kinect.

Furthermore, concentrical rings can be observed within the gray-coded image in Figure 10a. These artifacts might be caused by inhomogeneous areal illumination by the photodiode arrays, which results in range measurement inaccuracies due to the varying signal-noise-ratio of the range measurement.



a

b

Figure 12. Density distribution (blue dotted) and mean (red solid) of depth z versus standard deviation of the depth difference σ_z: a) PMD [vision] CamCube 2.0, b) Microsoft Kinect.

The Microsoft Kinect is difficult for interpretation, as no systematic error can be detected. Furthermore, a low point density is given at depths above 19 m, which could be interpreted as limitation of the device. Concerning the scene contents, only the four nearest rows of chairs can be recognized

within the image in Figure 10b. This is even more clearly presented within the density distribution in Figure 11b, following the vertical direction. Concerning the reliable pixels, 18322 depth difference values (17.5%) are within the interval [-3,0] m. The mean value depicted in Figure 12b shows the standard deviation of the depth difference, which could be roughly generalized. Transferring this information, it could be interpreted that for instance at a depth of 10 m a measurement deviation of approximately 0.2 m can be expected and at 15 m a measurement deviation of approximately 0.5 m.

7. CONCLUSION AND OUTLOOK

In this paper, a semi-automatic approach for co-registration of data captured by range imaging devices with different configurations has been proposed. This allows for evaluating the absolute and relative accuracy of the range imaging devices. After registration, the depth difference and the standard deviation of the depth difference have been estimated for two range imaging devices, namely Microsoft Kinect and PMD [vision] CamCube 2.0.

Based on the established 3D-to-2D correspondences, the data captured with the test devices can be used to complete or densify sparse data captured with a reference device. Even more, the point clouds captured with both devices do not necessarily have to provide the same density or accuracy. Hence, the test devices provide additional information about the local environment even beyond the sensor specifications, e.g. the different rows of chairs can still be distinguished and the rough structure of the scene can be recognized. However, in this case, the measured 3D coordinates are significant less accurate for the Microsoft Kinect whereas for the PMD [vision] CamCube 2.0, hardware-based unwrapping procedures using different modulation frequencies yield a measurement accuracy which approximately remains on a constant and relatively low level.

Concerning the utilized data, it can be stated that the intensity of the test data derived from the Microsoft Kinect not always matches to the reference data, due to the different wavelengths of the devices. For a fully automatic approach, these different characteristics will cause that the automatic detection of the point correspondences will fail.

In contrast to this, test data derived from the PMD [vision] CamCube 2.0 matches sufficiently to the reference data. First investigations show that an automatic registration between the different data types can reliably be established via keypoint detectors, e.g. by using SIFT features (Lowe, 2004). However, it has to be mentioned that this device shows limitations arising from its image size, but it can be expected that this will be improved in close future.

Not yet investigated, another straightforward approach might be to consider the texture given by the range image instead of the above mentioned intensity image, because the geometric aspects are invariant to the utilized wavelength. However, the nearly monostatic configuration of the PMD [vision] CamCube 2.0 and the bistatic configuration of the Microsoft Kinect while capturing the data might lead to inconsistencies within the range image and this could be critical for processing.

The promising results of this paper show that the presented methodology has a high potential for automated co-registration of data captured with ranging devices which show different configurations concerning the measurement principle, point density and range accuracy.

ACKNOWLEDGEMENT

The authors would like to thank Eva Richter from the Geodetic Institute at KIT for assistance during the measurement campaign. Furthermore, we would like to thank Nicolas Burrus from Robotics Lab, University Carlos III, Spain for his open-source software package that helped us capturing the Microsoft Kinect data.

REFERENCES

Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24 (6), pp. 381-395.

Jutzi, B., 2009. Investigations on ambiguity unwrapping of range images. In: Bretar, F., Pierrot-Deseilligny, M., Vosselman, G. (Eds.). Laserscanning 2009. International Archives of Photogrammetry and Remote Sensing 38 (Part 3/W8), pp. 265-270.

Jutzi, B., 2011. Extending the range measurement capabilities of modulated range imaging devices by time-frequency-multiplexing. AVN - Allgemeine Vermessungs-Nachrichten.

Kahlmann, T., Remondino, F., Guillaume, S., 2007. Range imaging technology: new developments and applications for people identification and tracking. In: Beraldin, J.-A., Remondino, F., Shortis, M. R. (Eds.) Videometrics IX, SPIE Proceedings Vol. 6491, 64910C.

Konolige, K., Mihelich, P., 2010. Technical description of Kinect calibration, http://www.ros.org/wiki/kinect_calibration /technical, last access in May 2011.

Lange, R., 2000. 3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology. PhD thesis, University of Siegen.

Lepetit, V., Moreno-Noguer, F., Fua, P., 2009. EPnP: An accurate O(n) solution to the PnP problem. International Journal of Computer Vision 81 (2), pp. 155-166.

Lichti, D. D., 2008. Self-Calibration of a 3D Range Camera. International Archives of Photogrammetry, Remote Sensing and Spatial Geoinformation Sciences 37 (Part B5), pp. 927-932.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60 (2), pp. 91-110.

Maas, H.-G., 1992. Robust automatic surface reconstruction with structured light. International Archives of Photogrammetry and Remote Sensing 29 (Part B5), pp. 709-713

Moreno-Noguer, F., Lepetit, V., Fua, P., 2007. Accurate noniterative O(n) solution to the PnP problem. IEEE 11th International Conference on Computer Vision, pp. 1-8.

Reulke, R., 2006. Combination of distance data with high resolution images. In: Maas, H.-G., Schneider, D. (Eds.) ISPRS Commission V Symposium: Image Engineering and Vision Metrology, International Archives of Photogrammetry, Remote Sensing and Spatial Geoinformation Sciences 36 (Part B).

Salvi, J., Pages, J., Batlle, J., 2004. Pattern codification strategies in structured light systems. Pattern Recognition 37 (4), pp. 827-849.

Shan, J., Toth, C.-K., 2008. Topographic Laser Ranging and Scanning: Principles and Processing. Boca Raton, FL: Taylor & Francis.

Vosselman, G., Maas, H.-G., 2010. Airborne and Terrestrial Laser Scanning. Whittles Publishing, Caithness, Scotland, UK.

Weinmann, Ma., Weinmann, Mi., Hinz, S., Jutzi, B., 2011. Fast and automatic image-based registration of TLS data. ISPRS Journal of Photogrammetry and Remote Sensing.

STITCHING LARGE MAPS FROM VIDEOS TAKEN BY A CAMERA MOVING CLOSE OVER A PLANE USING HOMOGRAPHY DECOMPOSITION

E. Michaelsen

Fraunhofer IOSB, Gutleuthausstrasse 1, 76275 Ettlingen, Germany
eckart.michaelsen@iosb.fraunhofer.de

KEY WORDS: Panorama stitching, planar homographies, homography decomposition, underwater mapping, UAV surveillance

ABSTRACT:

For applications such as underwater monitoring a platform with a camera will be moving close to a large roughly planar scene. The idea to map the scene by stitching a panorama using planar homographies is nearby. However, serious problems occur with drift caused by uncertainty in the estimation of the matrices and un-modelled lens distortions. Sooner or later image points will be mapped to infinity. Instead this contribution recommends using the homographies only for the composition of local patches. Then the homography obtained between the first and the last frame in such patch can be decomposed giving an estimate of the surface normal. Thus the patches can be rectified and finally stitched into a global panorama using only shift in *x* and *y*. The paper reports about experiments carried out preliminarily with a video taken on dry ground but a first under water video has also been processed.

1. INTRODUCTION

1.1 Intended Applications

In particular underwater robot vision is restricted to keep the distance between a structure to be monitored and the platform on which the camera is mounted short. There are ideas to enlarge the allowable distance by using gated viewing devices [9], but in the waters found where the application is supposed to be located there will always be a maximal distance where no considerable image quality is allowed due to floating obfuscation. Good image quality can often be expected from imagery taken at distances such as one meter. On the other hand the structure to be mapped may well have an extension of several hundred meters. Here we restrict ourselves to roughly and locally planar structures – such as retaining walls, harbour structures or underwater biotopes. The goal is to stitch a kind of orthophoto from a long video sequence.
Under water the drift problem – as outlined in section 1.2 - is very serious. But it also occurs in unmanned aerial vehicle mapping, where the platform may be cruising in about a hundred meter height over a roughly and locally planar world of much larger extension, e.g. several kilometres in extension.

1.2 Problem

The standard state-of-the-art method for stitching of an image sequence into a larger panorama is driven by successive planar homography estimation from image to image correspondences between interest points. Most often it is assumed tacitly or explicitly that the camera should only rotate round the input pupil and not move around in space. If the scene is strictly planar, there is – in principle - no difference between the image obtained by a wide-angle view from close up (in pin-hole geometry and taken normally) and a view from further away, or even an ortho-normal map. So the stitching of large views using homographies should be equivalent to taking an ortho-normal map.
However, deviations from the planar scene form, e.g. when a retaining wall is only locally planar but cylindrical in its global shape, cannot be treated this way. Moreover, if the first frame of the video sets the reference – as is often done – it may well not be exactly normal to the scene. Then there exists a distance in which the plane through the camera location and normal to its focal axis will intersect the scene plane at a line somewhere. Points on this line will be mapped to infinity if the homography estimation were precise – and points beyond this line would appear on the opposite end of the panorama. If we are only one meter away from a structure of hundreds of meters this is to be expected.
More seriously, the homography sequence approach accumulates the inevitable errors in large chains of matrix multiplications. Such drift may contain un-biased parts from uncertainty in the interest point locations, but it also may contain biased parts. E.g. homography estimation tends to hide un-modelled lens distortions in the rotational part of the homography [2].

1.3 Related Work

Many panorama stitching software packages are commercially available or can be downloaded for free from the web such as HUGIN [1]. The theory of optimal estimation of homogenous entities, such as planar homography matrices, with more entries than degrees of freedom from image to image correspondences with proper uncertainty propagation has reached a high level of sophistication [5]. RANSAC methods for robust estimation of such entities are standard today [4,7] but there are also alternatives such as iterative reweighting or GoodSAC [11]. Under water panorama stitching has been addressed e.g. by [2] with particular emphasis on the lens distortion induced drift.

2. STITCHING LOCAL PANNOS INTO A LARGE MAP

2.1 Homography Estimation

A planar homography is a mapping $x'=h(x)$ from one image into the other keeping straight lines straight. Here x and x' respectively are the points in the images. Homographies form an

algebraic group with the identity as one-element. Using homogenous coordinates the homographies turn out to be linear: $x'=Hx$. Where H is a 3x3 matrix whose entries depend on the choice of the coordinate system in the images. This linear description hides the highly non-linear nature of homographies in the division when transforming x' back into inhomgenous image coordinates. Thus homographies may map a finite point into infinity and they are not invariant for statistically important entities such as centre of gravity or normal distributions. Still, there is consensus today that homographies can be estimated from a set of four or more correspondences of interest points using the linear matrix equation provided that 1) a coordinate system is used that balances the entries into the equation system such that signs have equal frequencies and absolutes are close to unity [7], and 2) H is not too far away from the unity matrix (in particular the "projective" entries H_{31} and H_{32} should be small).
In a video sequence 1) can be forced and 2) can be assumed. Thus, we follow the usual procedure using an interest points, correlation for correspondence inspection, and RANSAC [4] as robust estimator. The activity diagram in figure 1) gives the details of the procedure. In each frame of the video a set of interest points $\{p_{in}\ ;\ i=1,\ \dots\ ,\ k_n\}$ is extracted using the well-known squared averaged gradient operator in its Köthe variant [6,8]. These are tracked back in the previous frame also using standard functions – here optical flow including image pyramids from open CV base [12]. Among these a consensus set is selected and simultaneously an optimal homagraphy using linear estimation and RANSAC on the correspondences of the p_{in} in coordinates transformed accordingly [7].

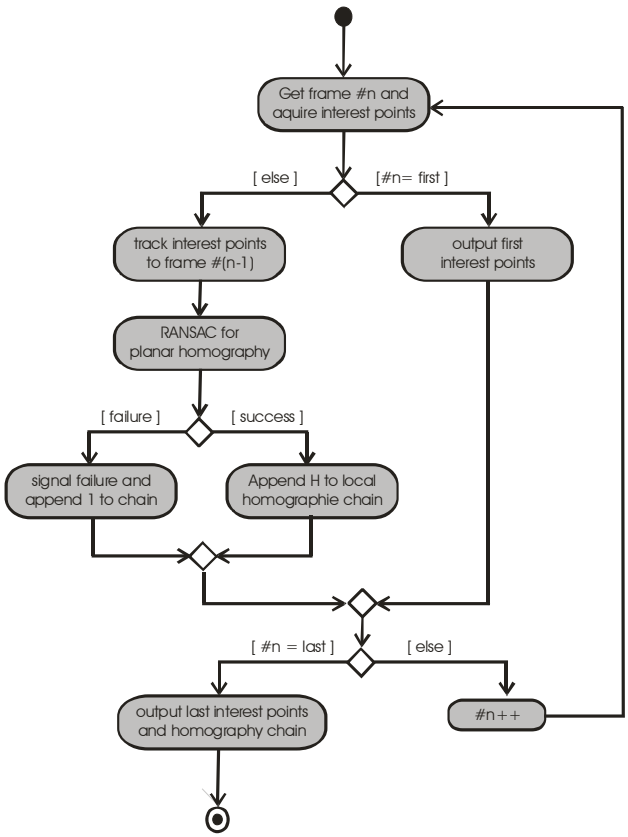


Figure 1. Activity diagram for partial homography chain estimation

Homographies cannot only be estimated for successive video frames but also for frames further apart from each other as long as there is sufficient overlap. However, if there is no sufficient overlap anymore the homographies must be chained in a sequence – by successive multiplication of matrices. Since there is uncertainty in the entries of this product there will be a drift – also in the "projective" entries H_{31} and H_{32}. Sooner or later points from an image far away from the first frame will thus be mapped to infinity.

2.2 Homography Decomposition and Rectification

Here H must be given in the normalized form, i.e. with the image coordinate system transformed such that the focal length equals unity and the principle point of the camera equals the origin of the coordinate system. So focal length and principle point should be known in good approximation. The standard decomposition of the matrix H in the form

$$H = \lambda R + tn^T \tag{1}$$

is known since [3]. Here R is the rotation matrix of the camera between the images, t is a translation vector, n is the surface normal of the planar scene, and λ a scalar factor. t can also be interpreted as homogenous entity. Than it is the image of the other camera, the epipole. n can also be interpreted as homogenous line equation. Than it is the line at infinity or the horizon of the scene.
This is the most important result here. The application demands that a proper – close to orthonormal – mapping of the scene should yield $n_1=n_2=0$ i.e. the normal identical to the viewing direction. After decomposition of H this can be achieved by applying appropriate rotations round the x and y axes:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix} \text{ and } \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix} \tag{2}$$

where $\beta=atan(n_1/n_3)$ and $\alpha=atan(n_2/n_3)$ after the rotation round the y axis. With this transformation the view should be rectified. We refer to [10] for a detailed analysis of the decomposition. There also a purely analytical solution to the decomposition can be found using only roots. Here the classical singular value decomposition version is used decomposing H into a product $H=UDV$. The entries of the central diagonal matrix d_{11}, d_{22}, d_{33} are the critical parts. They must be of sufficiently different sizes. Their differences are used as denominator while solving the quadratic equation system.
Two significantly different solutions appear among which we pick the one with n closest to $(0,0,1)^T$. The other solutions are flipped sign versions of no interest. But if the two solutions are equally close to $(0,0,1)^T$ or if the singular values are too similar the decomposition fails (resulting in a failure branch in the flow in Figure 2).

2.3 Stitching the Local Patches into a Large Panorama

It is our intention to treat all rectified panorama patches equal. Neither any projective distortion should be applied to them anymore – since this was corrected by homography estimation, decomposition and rectification, nor any shearing – since this is excluded by sensor construction, nor any scaling – since we assume that the platform is capable of sensing, controlling and keeping its distance to the scene plane. The rotations round the x and y axes where fixed in the rectification step. We will also assume that the camera is not rotating round the z axis on the long run by means of appropriate other sensors on-board the

platform – e.g. gravity sensor under water or compass on a UAV. The only two remaining degrees of freedom are shift in *x* and *y* direction.

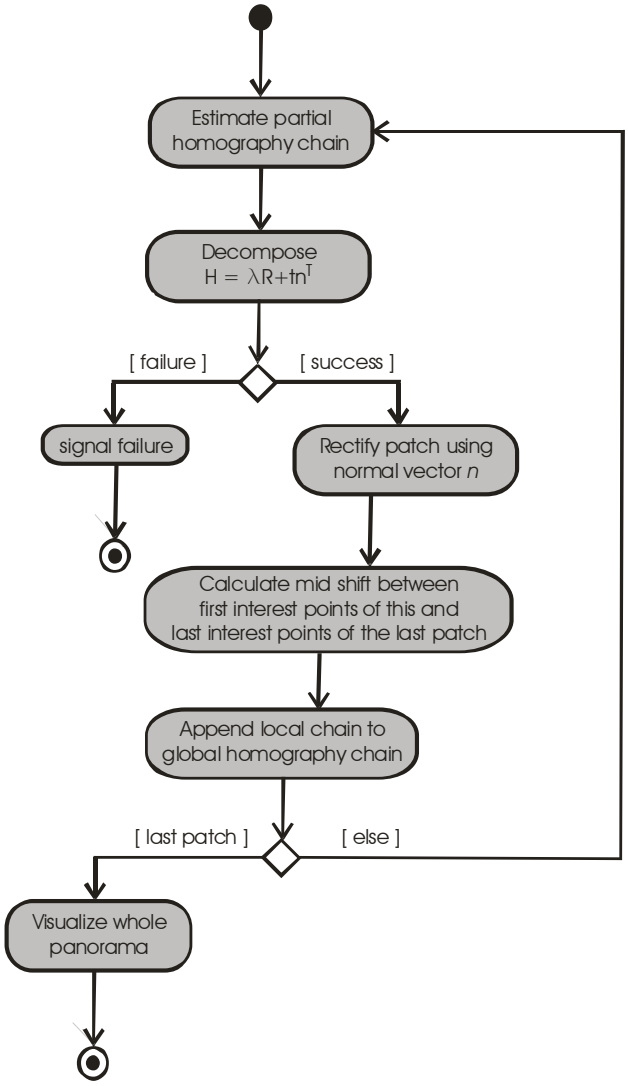


Figure 2. Activity diagram for global homography chain composition

This translation can easily be obtained by averaging the shift between the $p_{i,last}$ and $p_{i,first}$ of two successive patches. Recall, that the first image of a new patch is identical with the last image of the previous patch. Running the interest operator with the same parameters on the same image will give the same number of interest points in the same sequence. Such algorithms are deterministic. $p_{i,last}$ and $p_{i,first}$ of two successive patches are subject to different homographies, $p_{i,last}$ as result of a chained homography estimation plus a rectification and $p_{i,lfirst}$ as result of only a rectification. So there will be a residuum in this averaging process, which quantifies the success of the approach. But there cannot be any outliers.

Again a UML activity diagram gives an overview over this procedure (Figure 2). Here the stitching of a local panorama patch – i.e. the estimation of a partial homography chain as given in Figure 1 is hidden in one node.
This is still dead reckoning – since there is a possibly long sum of successive vectors with uncertainty drift – but it is much more stable than the multiplicative drift of the matrix chain. It is impossible that image points will ever be mapped to infinity

2.4 Resampling a Panorama from a Video

The main output of both a local estimation for patches as well as the global estimation for a panorama is a chain of homographies. So for each frame *i* of the video there is a homography h_i mapping a location – i.e. a line- and column index - of the panorama $(l_p , c_p)^T$ to a location in the *i*-th video frame $(l_{vi} , c_{vi})^T = h_i(l_p , c_p)^T$. However, a homography is a function mapping continuous coordinates into continuous coordinates. So if the panorama has similar or higher resolution than the video some type of interpolation will be required in order to fill the panorama with gray-values or colours from the video. Here the panorama is usually of lower resolution. So the coordinates in the video frame can be obtained simply by rounding $(l_{vi} , c_{vi})^T$.

Moreover, several frames of the video may contribute to the gray-value or colour to be displayed in one panorama pixel. The following possibilities are discussed:

- Averaging the value from all accessible frame locations $\{(l_{vi} , c_{vi})^T; 1 \le l_{vi} \le l_{max} \text{ and } 1 \le c_{vi} \le c_{max}\}$. This treats all information equally, but may give fuzzy results.
- Maximizing the gray-value over the index *i*. This is fast and easy, because all the non accessible positions either yield zero or NAN, but it has a bias towards brighter areas.
- Minimizing the distance to the centre $(l_c , c_c)^T$ using any metric $d_i = d((l_{vi} , c_{vi})^T , (l_c , c_c)^T)$ picks the gray-value from one particular frame. Here faults in the estimation may show up as sharp edges. We used this option here in order to explicitly show such problems.
- Maximizing the probability of a gray-value or colour given a drift model for the homographies and the measurements in the images (l_{vi} , c_{vi}) [13]. This needs assumptions on the uncertainty (e.g. normal distribution) and estimation of the parameters. Essentially, it leads to weighted averaging giving higher weight to gray-values from the centre. This needs most computational effort and diligence in parameter estimation – but leads to best and seamless results.

3. EXPERIMENTS

Some experiments where done outside of the water with a video taken by an Olympus PEN E-P1 camera with the standard zoom lens set to the extreme wide angle f=14mm. At this setting the lens shows considerable distortions giving slightly bending wall grooves (see Figure 3). This was not calibrated or modelled. The camera moves in a distance of about 0.7 meter along a wall constructed from large roughly axed stones. The scene is roughly planar with deviations of about three centimetres. The camera was kept mostly normal to the surface – but free handed. This fairly well mimics the kind of videos that could be expected from an underwater vehicle cruising along a retaining wall. On the other hand, outside of the water we can easily step back and take a groundtruth picture with a longer focal length and less distortion. The one presented in Figure 6 was taken with a Pentax istD S using a standard SMC 1:2 f=35mm lens. Still this is not calibrated – however it is sufficiently free of distortions since this is not a zoom lens, and it can also be used

on the larger 35mm film frame. Moreover, only a section from the image centre is used.

Figure 3. One frame of the test video

Figure 4. Panorama patch from a hundred frames using standard homography estimation

Figure 5. Rectified panorama patch using homography decomposition following [3]

From the HD video taken with the Olympus PEN local panorama patches were stitched using the standard flow outlined in section 2.1 (Figure 1). Rather arbitrarily we set the number of frames to be composed into one patch to one hundred. One such patch can be seen in Figure 4. While the view seems fairly normal on the left hand side – where the initializing frame was – it is evident that the projective drift effects already start at the right hand side of the patch (the moment that no overlap is given). We can see the kind of problem homography stitching has by using our knowledge that the stones are truly rectangular – see groundtruth in the upper picture of Figure 6. A certain drift – in particular in the "projective" entries H_{31} and H_{32} is inevitable. Figure 3 shows the rectification of this patch using the decomposition method described in Section 2.2 on the homography corresponding to the patch. It finds a reasonable compromise correcting the mistakes. In particular the rectangular structure is reproduced better. Some shear drift remains. This rectified patch is than part of the larger panorama displayed as lower picture in Figure 6, which was obtained by the method indicated in Section 2.3. It can be seen that a beginning drift is sometimes corrected by force introducing considerable non-continuous steps into the homography chain.

4. CONCLUSION AND OUTLOOK

Here we could only present a very preliminary overview of the intended system. It was mainly tested on videos from outside water with mild distortions and rather good quality. Less favourable data can be expected from under water platforms. On such data often nothing can be seen. If something is seen the lighting may well be quite inhomogeneous, there will probably be floating clutter in front of the interesting scene, and the lens may be out of focus – autofocus should be off in order not to be disturbed by the clutter. Lately, we obtained such a video, and one frame of it is presented in Figure 7.

The processing chain as indicated in the activity diagrams above hast to be adjusted to such situations. The same parameter values, e.g. for the interest operator, as applied to the test sequence of section 3 (just the usual default settings) give less than four interest points sometimes and on many other occasions RANSAC still fails to come up with a plausible solution. The computational flow only took the "else" path of the partial homography chain diagram once in more than thousand images of the example video, while persistently staying on that side for hundreds of (successive) frames of the

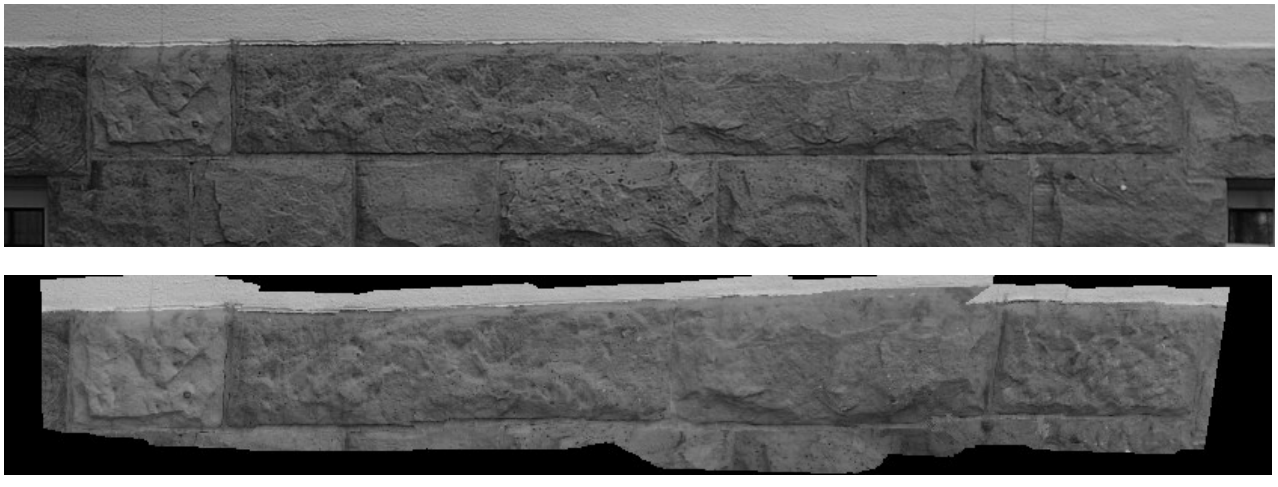


Figure 6: Lower: A panorama stitched from 5 patches, i.e. 500 frames, leading far away from the original frame on the leftmost side; Upper: Groundtruth picture of the same scene taken from further away

underwater video with the default parameters. Leaving the default settings in direction to more liberal ones of course gives less stable behaviour of the whole thing. Still, a preliminary result displayed in Figure 8 indicates sufficient stability to cope with such data. It is a good advice for underwater inspection to steer the vehicle as close to the structure of interest as possible.
It also becomes evident that the projective drift problem occurring when large sequences of such close-up videos are stitched can be mitigated by allowing full homography only on a local scale and keeping the global transform fixed to simple 2D-translation. The decomposition of the homography between the first and last frame of a patch giving an estimate of the surface orientation of the scene turns out to be an important help for rectification and subsequent joining of the patches.

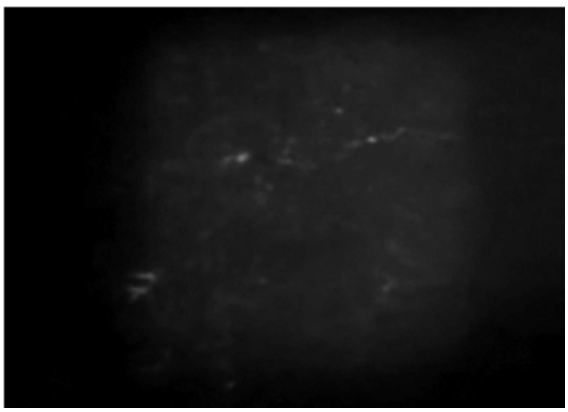


Figure 7: One frame of a typical underwater video

Obviously there is a trade-off between large patches from long camera movements allowing a stable decomposition with little error on the surface normal and epipole estimation on the one hand and the indicated projective drift problems that immediately begin to occur when there is no sufficient overlap anymore. Setting this parameter to a hundred frames can only be a first guess that has to be replaced by a mathematical investigation searching for the optimal patch size. Of course we look forward to making more experiments with challenging under water videos in the future. There remains a lot of room for improvement in all steps of the method.

References

[1] d'Angelo, P. HUGIN 2010.4.0, free panorama stitcher, http://hugin.sourceforge.net/download/ (accessed 28 Apr. 2011)

[2] Elibol, A., Moeller, B., Garcia, R., October 2008. Perspectives of Auto-Correcting Lens Distortions in Mosaic-Based Underwater Navigation. *Proc. of 23rd IEEE Int. Symposium on Computer and Information Sciences (ISCIS '08)*, Istanbul, Turkey, pp. 1-6.

[3] Faugeras, O., Lustman, F., 1988. Motion and structure from motion in a piecewise planar environment. *International Journal of Pattern Recognition and Artificial Intelligence*, 2(3), pp. 485–508.

[4] Fischler, M. A., Bolles, R. C., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the Association for Computing Machinery,* 24(6), pp. 381–395.

[5] Foerstner, W., 2010. Minimal Representations for Uncertainty and Estimation in Projective Spaces. *ACCV* (2), pp. 619-632

[6] Harris, C., Stephens, M., 1988. A combined corner and edge detector. *Proceedings of the 4th Alvey Vision Conference*. pp. 147–151. http://www.bmva.org/bmvc/1988/avc-88-023.pdf

[7] Hartley, R., Zisserman, A., 2000. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge.

[8] Köthe, U., 2003. Edge and Junction Detection with an Improved Structure Tensor. In: Michaelis, B., Krell, G. (Eds.): *Pattern Recognition, Proceedings* 25th*-DAGM*, Springer LNCS 2781, Berlin, pp. 25-32

[9] Laser Optronics, Underwater Gated Viewing Cameras, http://www.laseroptronix.se/gated/aqly.html, (accessed 28 Apr. 2011)

[10] Malis, E., Vargas M., Sep. 2007. Deeper understanding of the homography decomposition for vision-based control. INRIA report no. 6303, Sophia Antiopolis, France. http://hal.inria.fr/docs/00/17/47/39/PDF/RR-6303.pdf (accessed 28 Apr. 2011)

[11] Michaelsen, E., von Hansen, W., Kirchof, M., Meidow, J., Stilla, U., 2006. Estimating the Essential Matrix: GOODSAC versus RANSAC. In: Foerstner, W., Steffen, R. (eds) Proceedings Photogrammetric Computer Vision and Image Analysis. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Science,* Vol. XXXVI Part 3.

[12] Open CV Sources and DLLs, in particular http://opencv.willowgarage.com/documentation/cpp/video_motion_analysis_and_object_tracking.html?highlight=opticalflow#calcOpticalFlowPyrLK (accessed 26 Jun. 2011)

[13] Ren, Y., Chua, C.-S., Ho, Y.-H., 2003. Statistical background modeling for non-stationary camera. Pattern Recognition Letters (24), pp. 183–196

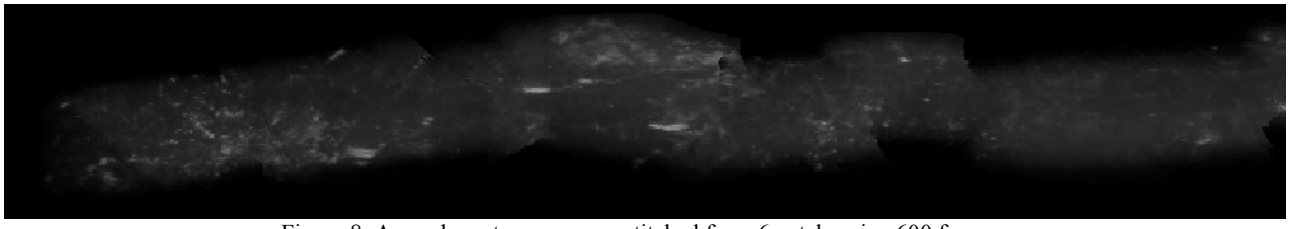


Figure 8: An underwater panorama stitched from 6 patches, i.e. 600 frames

WINDOW DETECTION IN SPARSE POINT CLOUDS USING INDOOR POINTS

S. Tuttas, U. Stilla

Photogrammetry and Remote Sensing, Technische Universität München, 80290 München, Germany - sebastian.tuttas@bv.tum.de, stilla@tum.de

KEY WORDS: point cloud, building, city, façade reconstruction, airborne laser scanning

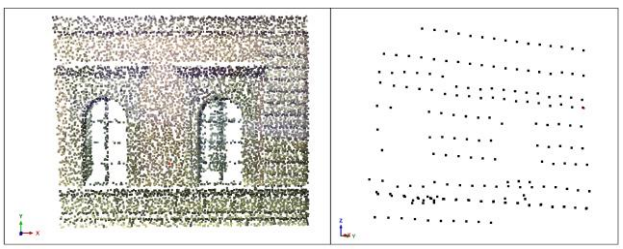
ABSTRACT:

This paper describes an approach for detecting windows from multi-aspect airborne laser scanning point clouds which were recorded in a forward looking view. Since the resolution of the point cloud is much lower than from terrestrial laser scanning, new methods have to be developed to detect and, in a further step, reconstruct façade structures. The façade planes are detected using point normals and a regiongrowing algorithm. The approach for window detection uses the points which are lying behind the detected façades planes (indoor points). Regularities in the appearance of these points are of special interest to enable the detection of windows which are only weakly represented in the point cloud. Therefore it is checked with a Fourier Transform if a repetitive structure can be extracted. Otherwise peaks in the density of the indoor points are used to detect the windows. The approach is tested on data from four overflights over the area around the TU München. The tests show that windows having a repetitive structure can be detected well for larger façade parts which provide enough samples but the approach shows deficits for small façade parts and in the case of disturbing intrusions.

1. INTRODUCTION

1.1 Motivation

3D-city models are used for several applications, for example urban planning, navigation or visualisation of buildings of touristic interest. In these cases it is sufficient to use polyhedral models. They can be used with or without texture. But there are also applications which require a more detailed façade reconstruction. To analyse Persistent Scatterers for radar image interpretation façade details are of special interest (Auer et al. 2011), for energetic assessment of buildings with thermal cameras the area of the façade without the window regions is needed (Iwaszczuk et al. 2011). Also the visualisation of a building can be made more realistic by modelling geometric structures like windows and doors.
In most cases the polyhedral models are received from airborne laser scanning data. From this the roof structure can be modelled but only less points can be found on the façade.



a) b)

Figure 1. Example for laser point clouds: a) Dense point cloud (~ 200 points / m^2) from terrestrial laser scanning. b) sparse point cloud (~ 5 points / m^2) from forward looking airborne laser scanning

Façade reconstruction methods make usually use of terrestrial laser scanning data, e.g. from a street mapper. This data has often a very high point density, but lacks of roof data. Normally only the part of the façade can be seen, which is oriented towards the street.
Airborne multi-aspect laser scanning data makes it possible to reconstruct building façades and roofs for entire buildings from a single data set. This data is a compromise between completeness and point density, which cannot be as high as from terrestrial scanners. In Figure 1 the point density of a terrestrial point cloud and from our test data is compared.

1.2 Related Work

A comprehensive overview on 3D building reconstruction from LiDAR and from image data is given by Haala and Kada (2010). Their paper has two main parts, one describes approaches for roof shape reconstruction, the other one outlines approaches on building façades. They state a large variety of different works on the topic of roof shape reconstruction, whereby they distinguish between three main groups: reconstruction with parametric shapes, reconstruction based on segmentation and reconstruction by DSM simplification.
Since the roof reconstruction is more advanced and the different approaches are well described in Haala and Kada (2010) only papers concentrating on façade reconstruction are presented in the following.
A basic problem is always the detection of structures in the point cloud, mainly planes. An overview on this topic is given by Vosselmann et al. (2004).
Ripperda (2008) derived grammar rules for façade parameters from images which can be used in the reconstruction process using a Reversible Jump Markov Chain Monte Carlo method. Also Becker (2009) uses a formal grammar to reconstruct building façades. The grammar is derived from terrestrial laser scanning data and is refined with image data. With the help of the grammar also building parts, which are occluded, can be modelled.
Boulaassal et al. (2009) detect window contours using a 2D-Triangulation of the façade plane. Since the windows are represented by holes in the façade, it is searched for the longest triangle sides to find the points surrounding the windows.

Schmittwilken and Plümer (2010) use a model-based reconstruction approach. They use training data to create probability density functions for the shape parameters of windows, doors and stairs which are used for a prefiltering of the point cloud. The selection of the most likely sample for a certain object structure is done by an adapted RANSAC approach which uses a more efficient criterion for the scoring.
Pu and Vosselman (2009) extract features from a segmented point cloud by defining constraints for the different façade features. They also use a hole-based window extraction method using a TIN. Knowledge is brought in to complete the parts of the building, which are occluded.

1.3 Concept

The approaches mentioned before have all in common that they use high resolution point clouds (hundred to several hundred points per square meter). Our data has only approximately 5 points per square meter, but there can also be parts with a total lack of points. Because of the oblique view the point density on the façade can vary for a single scan. This makes it hard to use holes in the point data for window detection, what lead us to an approach which uses regular patterns of points behind the façade (= indoor points). The proposed workflow can be found in Figure 2.

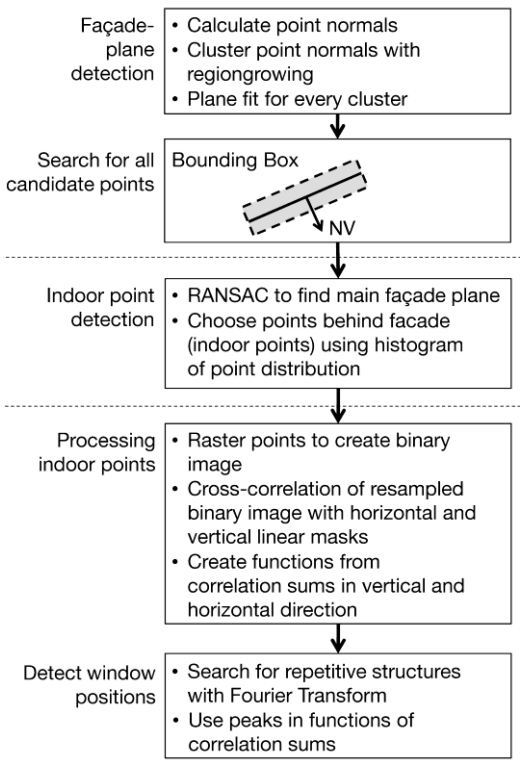


Figure 2. Flowchart of the presented approach

The approach is based on two basic assumptions:
- Laser pulses can pass through windows and are reflected inside the building. These points are a few cm to a few m behind the main façade plane.
- Windows are often arranged in a regular way, at least in one façade direction (vertical or horizontal). This also means that it is likely that a window exists at the same position on each floor.

First façade planes have to be detected, what is described in Section 2. After that every façade is processed on its own to detect the windows, what is delineated in Section 3. This has three parts. The first one is the detection of the indoor points (3.1), what is done by fitting a Gaussian function to a histogram of the point distribution. Then the indoor points are rastered to generate a binary image. This is cross-correlated with a horizontal and a vertical line (3.2). Finally it is searched for repetitive structures in the resulting correlation images (3.3). The approach is tested on a data set of the TU München, what is shown in Section 4. Conclusions and outlook are given in the last section.

2. FAÇADE PLANE DETECTION

First normals for every point have to be calculated using a search radius r depending on the point cloud density (e.g. 3 m). The normals are used to find potential façade points. All points having normals which deviate more than $\Delta\varphi$ from the horizontal plane are rejected. This helps to reduce the points, which have to be processed. The point normals can be used in different ways to support the segmentation process. For example Awwad et al. (2010) improved a RANSAC algorithm by including a check between the normal vector of the point cloud and the hypothesised RANSAC plane. Here a regiongrowing algorithm using the normals is applied to extract the façade planes from the point cloud (see Figure 3). For every point always the n nearest neighbours are considered. The points are allocated to the same segment if the distance between the points is less than Δd (distance threshold) and if the angle between the normal vectors, projected into the horizontal plane, is less than $\Delta\delta$ (angle threshold).

```
1:  seg_k =1
2:  for i = 1 : # points
3:    get n Nearest Neigbours (NN)
4:    calculate distance d and angle δ between normal vectors
5:    if d < Δd && δ < Δδ)
6:      find best NN (NN_best)
7:      if seg(NN_best)=∅ || seg(seg(P_i) )=∅
8:        if seg(P_i)=x: seg(NN_best)=x
9:        elseif: seg(NN_best)=y: seg(P_i)=y
10:       else: seg(NN_best)=seg(P_i)=k, k=k+1, seg_k+1=seg_k +1
11:     else
12:       all P with seg(P)=seg(NN_best): seg(P)=seg(P_i)
13:       if exists best NN with seg(NN)=∅
14:         seg(bestNN_∅)=seg(P_i)
15:   else
16:     if seg(P_i)=x: seg(P_i) = x
17:     else: seg(P_i)=seg_k, k=k+1 ; seg_k+1=seg_k+1
18: end
```

Figure 3: Proposed algorithm for façade plane detection

A plane is accepted if it is composed of at least j points, what is again dependent on the point density. As can be seen in the pseudo code the best fitting point has priority. The best fitting point is the point having the minimum product of distance and angle of normal vectors (in radian). Only this point is added to the segment of the processed point, of course only if the both thresholds are not exceeded. If the best fitting point and the processed point are allocated to different segments, these segments are fused and the best point, which has no segment yet, is also added to this segment. For every segment a plane is fitted using principal component analysis performed on the matrix with the points of the plane. The normal vector is the

principal component with the smallest covariance (Klasing et al. 2009). Subsequently the plane is forced to be vertical by projecting the normal vector into the x-y-plane. The planes are intersected to get the vertices of the building. This step is done manually yet, but shall be automated in future.
Not all the points before and behind the façade planes (indoor points, protrusions, intrusions) are included in the segmented façade points because they have not been regarded during the extraction of façade candidates or the regiongrowing process. Because of this there is a step back to the complete point cloud. All points are chosen which are in the range of *k* m (depends on maximum window height and looking angle of the laser) before and behind the plane. These are finally used for the window detection. From now all façades are processed on their own.

3. WINDOW DETECTION

3.1 Extracting indoor points

First the main façade plane, which is used as reference for the decision if a point lies on or behind it, is determined using a RANSAC algorithm. It is assumed that the plain with the most inliers from the points derived by the bounding box extraction should be the best reference plane.
The points are transformed into a coordinate system with the x-axis being orthogonal to the plane. A histogram is calculated showing the amount of points in a certain distance from the plane (see Figure 4).
Whether a point is declared to be lying behind the main façade plane or not is dependent on the façade roughness (flat surface or many protrusions/intrusions). Because of this a Gaussian function in the following form (see Eq. 1) is fitted to the histogram:

$$y = A\frac{1}{\sigma\sqrt{2\cdot\pi}}e^{\left(\frac{(x-\mu)^2}{2\sigma^2}\right)} \quad (1)$$

The position μ-σ is chosen as threshold for the indoor points (dashed line in Figure 4). The minimum threshold is defined as 10 cm. Finally the indoor points are projected into the façade plane regarding the incidence angle of the laser.

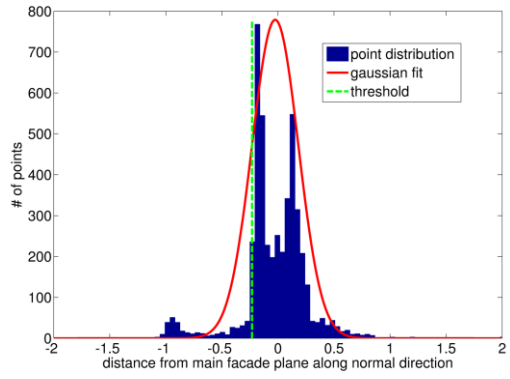


Figure 4. Example for a histogram of the point distribution along the normal direction of the façade, the red line is the fitted Gaussian function and the dashed line is the derived threshold to decide which points are indoor points.

3.2 Binary image and cross-correlation

A binary image with a resolution which is appropriate for the point density (e.g. 1 m) is created from the indoor points by setting a pixel to 1 if a point exists in the respective cell. The binary image is resampled, increasing the resolution with factor 10, and then cross-correlated with a template of a horizontal and a vertical linear mask of 2 m (an average window size) length (see Figure 5). The horizontal line is used to detect the window positions in x-direction (width) and the vertical line to detect the window positions in y-direction (height).

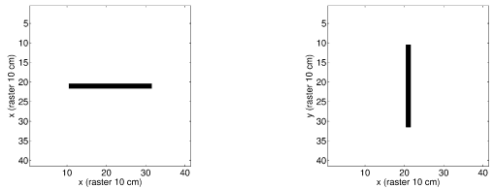


Figure 5. Horizontal and vertical line mask which is used for cross-correlation with the binary image.

Two functions are delineated from the correlation image. The entries of the correlation image are summed up in x-direction to get the positions of windows in y-direction and the other way round. One of these functions can be seen in Figure 6. The use of the line masks works like a smoothing, what leads to clearer peaks in the function of the sums.

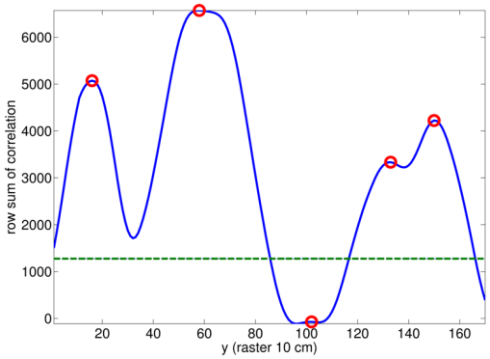


Figure 6. Example for a function of the sum over the values of all columns in the correlation image for each row (= height profile). The peaks in the function, indicating window positions, are marked with circles. The dashed line shows the threshold for accepting peaks as window positions.

3.3 Searching window positions

The search for the window positions is done independently for x- and y-direction. At the end a window is placed at every possible combination of x- and y-positions. The functions of the sums of the correlation images are used twice. First these signals are used as input for a discrete Fourier Transform to look for a repetitive structure. If no such structure can be found the peaks of the function are used as window positions.
As repetition frequency of the windows the best non-zero-frequency of the resulting spectrum is chosen. Three requirements khave to be fulfilled to accept the result of the Fourier Transform:

- The frequency has to lead to at least four windows for one façade plane in the respective direction. A signal generated

by fewer windows is too short to provide a reliable solution.
- The median of the distances from the peaks in the function and the frequency have to be equal in between a certain range.
- There has to be a significant peak in the spectrum.

If the result of the Fourier Transform is accepted, the windows are positioned over the whole façade with the determined frequency, starting from the position of the window with the highest correlation (highest peak in the function of the sums).
Windows lying too close to the edge of the façade (e.g. <1 m) are neglected.
If the frequency is not accepted, the positions of the peaks in the functions are used. Peaks which are below mean-peak-height/3 (dashed line in Figure 6) are neglected.
The following steps are carried out to improve the derived window positions:
- Peaks which are too close to the edge of the building are removed. This is necessary because there are still some points of the adjacent façades which are normally in front or behind the processed façade, what leads to clear peaks in the function.
- Peaks which are too close together are fused to one peak. The threshold is 1.5 m for the horizontal distance, and 2 m for the vertical distance, assuming that floors have at least a distance of 2 m.

4. EXPERIMENTS

4.1 Data

We use a dataset of the test area TUM (Technische Universität München) recorded by four overflights with a helicopter. The area was scanned in 45° oblique view, what leads to a point cloud, where all building façades can be seen from all directions. The co-registration of these four different point clouds can be done with homologous planes or an adapted ICP algorithm (Hebel and Stilla 2009, Hebel and Stilla 2007). The composed point cloud can be seen in Figure 7. The total points are around 2.5 million.

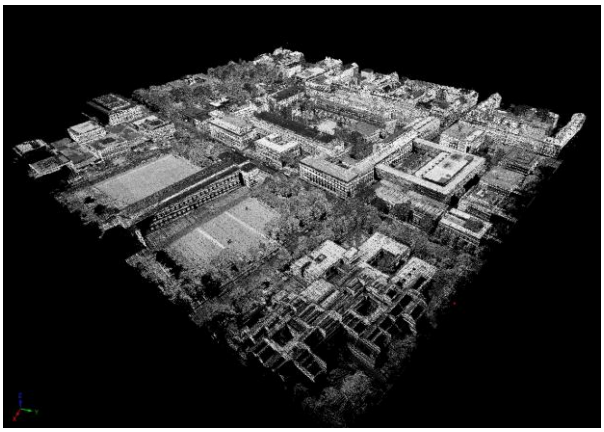


Figure 7. Point cloud with approximately 2.5 million points of TUM area composed of four point clouds from four overflights.

4.2 Façade plane detection

For façade plain detection a threshold of $\Delta\varphi = 10°$ is chosen for the coarse elimination of non-façade-points. Approximately 1/10 of the originally point cloud is remaining.
The algorithm shown in Figure 3 is run with $\Delta d = 3$ m, $\Delta\delta = 5°$ and $n = 10$. To reject small planes the threshold $j = 100$ for minimum points is used. The resulting segments can be seen in Figure 8.

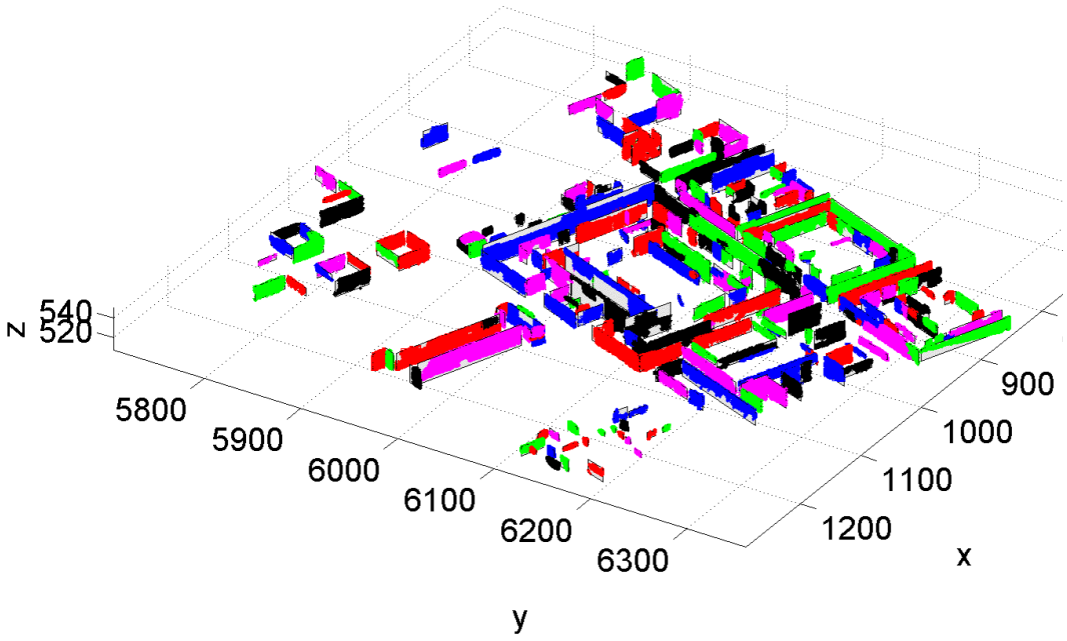


Figure 8. Segments with more than 100 points found by the regiongrowing algorithm

4.3 Window Detection

In the following the window detection is shown for one building (Old Pinakothek) of the test area. The selected points for the northern façade (see Figure 9) received from a bounding box using k = 5 m can be seen in Figure 10.

Figure 9. Northern façade of the Old Pinakothek

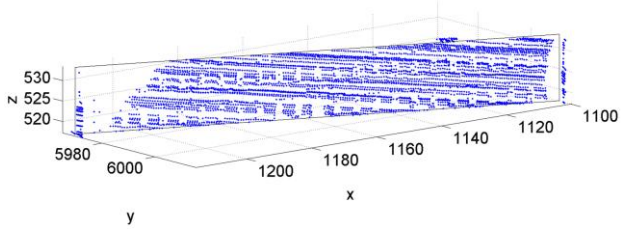


Figure 10. Points derived with a bounding box which delimits the points to the area of 5 m before and behind the façade (northern façade of the Old Pinakothek).

In Figure 11 the detected indoor points are shown. From these points the raster image in Figure 12 is computed.

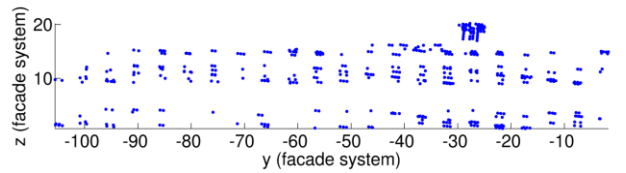


Figure 11. Indoor points derived from the points shown in Figure 10 and the threshold shown in Figure 4.

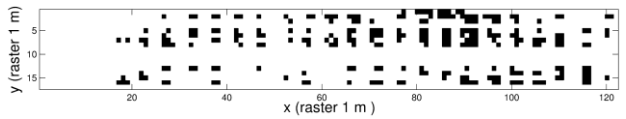


Figure 12. Binary image created by setting a pixel to 1 if any indoor point (shown in Figure 11) is inside the 1 m cell.

Finally in Figure 13 the detected windows can be seen for the façade shown in Figure 9.

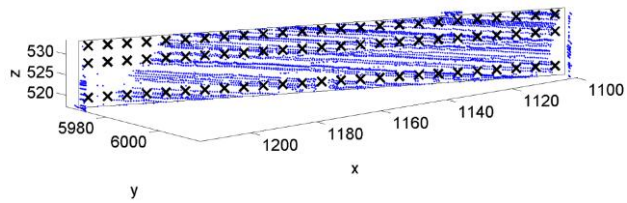


Figure 13. Façade points from Figure 10 representing the façade shown in Figure 9 with detected window (centre) positions.

In this case the number of windows in both directions was determined correctly. Since it is not possible to distinguish between doors and windows, the door in the middle of the façade is also marked as window. An appropriate frequency was detected for the windows along the width of the façade. For the window positions in height direction the peaks from the function in Figure 6 are used, whereby the last two peaks are fused.

4.4 Results

In Figure 14 ten façade planes of the Old Pinakothek with the detected window centre positions are shown. For evaluation the planes are divided into three groups: a) the two long planes, b) the front face and c) the seven small sides. In Tables 1 to 3 the evaluation results can be seen. Since there are doors in group a) and intrusions in group b) and c), which lie behind the main plane, these are also detected as windows. They are specified separately in the tables.

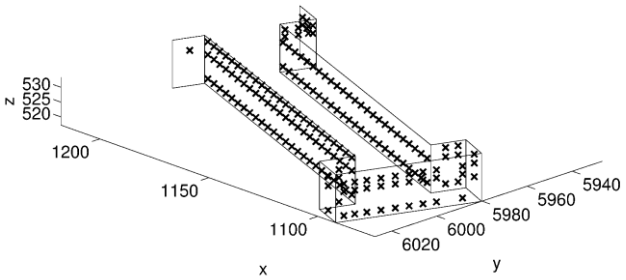


Figure 14. Ten façade planes of the Old Pinakothek, the black crosses represent the detected window positions

Group a)					
W	D	T	WD	DD	FA
123	2	125	123 (100%)	2 (100%)	-

Table 1. Result of the window detection for the two long planes, W = windows (ground truth), D = doors (ground truth), T = amount of detected windows, WD = correctly detected windows, DD = detected windows which are actually doors, FA = false alarm

Group b)					
W	I	T	WD	ID	FA
9	5	18)* 27	9 (100%)	5 (100%)	4)* (22%) 13 (48%)

Table 2. Result of the window detection for the front façade, W = windows (ground truth), I = intrusions (ground truth), T = amount of detected windows, WD = correctly detected windows, ID = detected windows which are actually intrusions, FA = false alarm;
)* these values are produced counting double detection of one (large) window/intrusion as one detection.

Group c)					
W	I	T	WD	ID	FA
7	7	17)* 28	3 (43%)	7 (100%)	7)* (41%) 18 (64%)

Table 3. Result of the window detection for the small façade parts, explanation of abbreviations see Table 2.

As can be seen from the results in Table 1 windows showing a regular pattern over a certain extent, like it is in group a), can be detected very well. For the front plane (b)) the problem occurs that there are intrusions which produce indoor points. These cannot be distinguished from indoor points, which are originated from real windows. Additionally this façade does not show a regular pattern over the whole plane.
The worst results can be found for group c). There is only one window, so the approach searching for a repetitive structure cannot work. The problem with intrusion also occurs here, what leads to many false alarms.

5. CONCLUSIONS AND FUTURE WORK

This paper has shown that a window sequence can be detected, if a signal can be produced by indoor points from a sufficient amount of regular arranged windows. From that follows that this approach is useful for urban scenarios, where often buildings with several floors and regular structure of windows can be found.
The result of the façade detection shows that it is useful to work with multi-aspect side looking airborne laser scanning. From that data façade planes at the back of the buildings, which cannot be seen from the street, or as in the case of the Old Pinakothek, façade planes which are too far away from the street to be acquired by a street mapper, can be provided.

Since the work on this approach is at the beginning, there are several possibilities to improve the results:

- The buildings have to be extracted automatically after the segmentation. This can be done using the normal direction of the façades which can be derived from the navigation data.
- It has to be found a way to distinguish between doors, windows and other intrusions. For that purpose other features like the point density in the façade plane or intensity values have to be considered.
- Separation of rows and columns to make it possible to detect different pattern for different rows/columns of one façade. This is especially important for the ground floor row, which often shows a special behaviour.
- The approach can be extended in a way that also protrusions can be detected.
- The window size shall be derived, for example by fitting a rectangle function in the function of correlation sums.
- The geometric accuracy of the detected window centres has to be evaluated.

ACKNOWLEDGEMENT

The authors want to thank Marcus Hebel (Fraunhofer IOSB) for providing the co-registrated point cloud.

REFERENCES

Auer S., Gernhardt S., Bamler R., 2011. Investigations on the nature of Persistent Scatterers based on simulation methods. In: *2011 Joint Urban Remote Sensing Event (JURSE)*, Munich, Germany, pp. 61-64.

Awwad T. M., Zhu Q., Du Z., Zhang Y., 2010. An improved segmentation approach for planar surfaces from unstructured 3D point clouds. *The Photogrammetric Record*, 25(129), pp 5-23.

Becker S., 2009. Generation and application of rules for quality dependent façade reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(6), pp. 640-653.

Boulaassal, H., Landes, T., Grussenmeyer, P., 2009. Automatic extraction of planar clusters and their contours on building façades recorded by terrestrial laser scanner. *International Journal of Architectural Computing*, 2009 (1), pp. 1-20.

Haala N., Kada M., 2010. An update on automatic 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6), pp. 570-580.

Hebel M., Stilla U., 2007. Automatic Registration of laser point clouds of urban areas. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Munich, Germany, Vol. XXXVI, Part 3/W49A, pp. 13-18.

Hebel M., Stilla U., 2009. Automatische Koregistrierung von ALS-Daten aus mehreren Schrägansichten städtischer Quartiere. *PFG Photogrammetrie, Fernerkundung, Geoinformation*, 2009 (3), pp. 261-275.

Iwaszczuk D., Hoegner L., Stilla U., 2011. Matching of 3D building models with IR images for texture extraction. In: *2011 Joint Urban Remote Sensing Event (JURSE)*, Munich, Germany, pp. 25-28.

Klasing K., Althoff D., Wollherr D., Buss M., 2009. Comparison of surface normal estimation methods for range sensing applications. In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA), 2009*, Kobe, Japan, pp. 3206-3211.

Pu S., Vosselman G., 2009. Knowledge based reconstruction of building models from terrestrial laser scanning data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(6), pp. 575-584.

Ripperda N., 2008. Determination of façade attributes for façade reconstruction. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Beijing, China, Vol. XXXVII, Part B3A, pp. 285-290.

Schmittwilken J., Plümer L., 2010. Model-based reconstruction and classification of façade parts in 3D point clouds. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Sainte-Mandé, France, Vol. XXXVIII, Part 3A, pp. 269-274.

Vosselmann G., Gorte B. G. H., Sithole G., Rabbani T., 2004. Recognising structure in laser scanner point clouds. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Freiburg, Germany, Vol. XXXVI, Part 8/W2, pp. 33-38.

INTERPRETATION OF 2D AND 3D BUILDING DETAILS ON FACADES AND ROOFS

P. Meixner [a, *], F. Leberl [a], M. Brédif [b]

[a] Institute for Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria
(meixner, leberl)@icg.tugraz.at
[b] Université Paris-Est, Laboratoire MATIS, IGN, 94165 Saint-Mandé, Cedex
bredif@ign.fr

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: Facades, roofs, semantic segmentation, aerial images, street side images, multi view

ABSTRACT:
Current Internet-inspired mapping data are in the form of street maps, orthophotos, 3D models or street-side images and serve to support mostly search and navigation. Yet the only mapping data that currently can really be searched are the street maps via their addresses and coordinates. The orthophotos, 3D models and street-side images represent predominantly "eye candy" with little added value to the Internet-user. We are interested in characterizing the elements of the urban space from imagery. In this paper we discuss the use of street side imagery and aerial imagery to develop descriptions of urban spaces, initially of building facades and roofs. We present methods (a) to segment facades using high-overlap street side facade images, (b) to map facades and facade details from vertical aerial images, and (c) to characterize roofs by their type and details, also from aerial photography. This paper describes a method of roof segmentation with the goal of assigning each roof to a specific architectural style. Questions of the use of the attic space, or the placement of solar panels, are of interest. It is of interest that roofs have recently been mapped using LiDAR point clouds. We demonstrate that aerial images are a useful and economical alternative to LiDAR for the characterization of building roofs, and that they also contain very valuable information about facades.

1. INTRODUCTION

Accurate and realistic 3-dimensional models of the urban human habitat are gaining importance for virtual tourism, city planning, internet search and many emerging municipal engineering tasks. They also represent location information for the evolving field of ambient intelligence. Internet search in Bing Maps or Google Earth is supported by 3D virtual cities worldwide. At this time these data are just used for visualization, but this is on the way to change. On the horizon are urban models that consist of semantically interpreted objects. In its most sophisticated form, each building, tree, street detail, bridge and water body is modeled in three dimensions, details such as windows, doors, facade elements, sidewalks, manholes, parking meters, suspended wires, street signs should exist as semantically identified objects.

Automatic mapping of facades and roofs in 3D is a fundamental element in building 3D virtual cities, but the tasks are surprisingly complex. We present in this paper several avenues of research we are pursuing to achieve automation from image sources, mostly in the form of aerial photography, but also in the form of street side images. For example, we have to define a building, its facades and roof with 3D detail extruding from or intruding into predominant planes. Such detail may be stair cases, balconies, awnings, dormers, chimneys, terraces, elevator shafts, air conditioning units, roof gardens and the likes.

Street side images are very important current sources of building information because such imagery is being created by both the vehicle-based industrial data collection systems as well as by Internet users in the form of Community Photo Collections. We therefore study facades using overlapping street level images. Recky et al. (2011) have shown that individual facades can be detected with a success rate of 97%. However, facades are also imaged in vertical aerial vertical image at the edge of the fields of view. Normally vertical aerial images are used for orthophotos and the mapping of roofscapes. The idea of using such data for the analysis of vertical walls may surprise. We demonstrate that the idea is valid, and that aerial vertical imagery is a good source to model building facades in 3D based on plane sweeps.

Mapping of roofs is also a fundamental element in characterizing buildings. The majority of research is based on LiDAR point clouds. We show that digital aerial images and point clouds extracted from them serve well for the characterization of building roofs. Roofs need to be modelled by their major planes and thus the architectural roof style, and by their 3D detail of chimneys, dormers, sky lights, terraces and such. An initial test area supports the conclusion that roof planes can be correctly mapped in 89% of all cases, and that the assignment of roofs to their roof type is successful at a rate of 82%.

* Corresponding author.

2. FACADE DETECTION USING STREET SIDE IMAGES

2.1 Approach

Street-side images need to get interpreted and facades need to get identified, each facade as a separate entity (Recky et. al. 2010, Hammoudi K. 2011). A street side will appear as a continuous agglomeration of connected buildings. At issue is the splitting of a building block into its individual buildings. The data source is a set of overlapping, thus redundant images taken from a moving vehicle carrying calibrated automated cameras.

An initial segmentation divides the image into different contents like sky, cloud, roof, building, ground, vegetation, shadow and undecided. The segmentation was described by Recky et. al. (2010) and computes image patches using a watershed segmentation. Patches are consequently merged into larger segments depending on color and texture. A graph is then constructed where every image segment is a node and the edges define the type of relationship of the segments with one another. These relationships are examined using discriminative random fields (DRF). The spatial relations between segments represent context and permit one to differentiate between ground and sky or roof and facade. In a test area in Austria, the detection of building facades achieved a success rate of 94%.

The result of the previous step produces facade areas, not individual facades per building. In a next step repeated patterns in the images get associated with separate facades. The approach was introduced by Wendel et al. (2010) based on Wendel (2009). The method uses Harris corners as interest points. In a next step the color profile between every interest point and the 30 nearest neighbors is calculated. The color profiles are constructed using a 20-dimensional normalized descriptor for each of the three colors RGB, in total thus with 60-dimensions. A kd-tree method is then used for matching the descriptors. In a last step the repetitive patterns are located in a voting matrix.

In a next step the processing of the single facade is discussed in more detail. Due to the natural settings of objects in these images we assume that repetitive patterns occur along the horizontal direction and the separation of the facades occur in vertical direction. Therefore the lines between the matched interest points are projected into the horizontal axis constructing a match cost histogram. Then the facades are segmented by determining a separation area (area where one facade ends and the next begins). This is done by defining areas with a low likelihood as separation areas and areas with high likelihood as repetitive areas. To be able to determine the exact split between two facades in a last step they look for the global maximum in these areas.

By applying the pattern-based facade separation on image segments previously identified as facade space, the results improve over those achieved without the use of facade spaces (Recky et al., 2011). Figure 1 illustrates the result of the facade segmentation showing 4 test images.

2.2 Experimental Results

Tests were based on 9 separate building facades shown in 20-50 overlapping photos. The images are taken in a forward look so that the facades are shown under an oblique angle. This helps in evaluating the influence of the perspective distortion. A detection rate was achieved of 97% (see Figure 1).

Figure 1: Four examples of automatically segmented street side images into individual building facades. Use is being made of overlapping street-side images. In this example, the camera is pointed forward (from Recky M. et al., 2011).

2.3 Discussion

The separation of buildings using their facades is an alternative to the separation of buildings by cadastral property data. Ideally, these two approaches produce the same result. However, buildings can extend over multiple properties, or one property can hold more than one building. Future work will have to address the comparison between the two alternatives and find means of resolving any discrepancies.

Future work will also have to deal with different architectural styles, non-plane facades, various imaging modalities and the problems with occlusions from vegetation. As we argue that aerial photography also offers facade information, the approaches thus far designed to work with street level images should be applied to aerial imagery as well.

3. FACADE DETECTION USING VERTICAL AERIAL IMAGES

3.1 Approach

Aerial vertical images are less affected by occlusions from vegetation and are available “freely” and at no cost of acquisition. The data will have been collected for traditional urban mapping anyway. Looking at facades in vertical aerial images offers therefore an added benefit. We combine aerial images with cadastral information. We have developed a framework for building characterization that is strictly built for aerial photography (Meixner et al., 2011). We start out by merging the aerial imagery with property boundaries to define each property as a separate entity for further analysis. The cadastral data may also contain preliminary information about a building footprint. In the next step the building footprints get refined vis-à-vis the mere cadastral prediction based on an image classification and on the definition of roof lines. 3D facade coordinates are computed from aerial image segments, the cadastral information and the DTM. This helps to determine the number of floors, the window locations (see figure 2) and offers candidates for attic and basement windows.

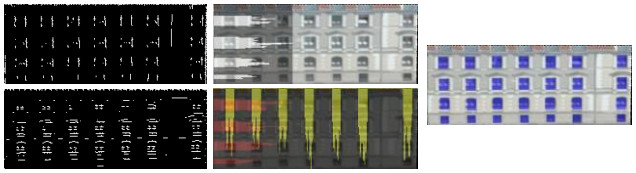


Figure 2: Processing steps for floor and window detection. (a) Horizontal and vertical edges, (b) maxima search in horizontal projection profile and overlay of the two profiles, (c) result of the window detection with highlighted window locations in blue. The count of floors and windows coincides with ground truth.

This method is well suited when a facade is generally plane, but fails with complex facades with extrusions like balconies, staircases and awnings. Figure 3 illustrates a failure.

Figure 3: To the left is a rectified facade image with a depth structure, to the right a failed count of windows. The 3D structure needs to get considered.

To be able to deal with those facades we have to reconstruct them in 3D so that we can separate a facade in planar segments. This is achieved with the so-called plane sweeping method along the proposal by Zach (2007), with its advantage that one no longer needs to assume a single vertical plane per facade but also complex facades with awnings, bay windows, staircases and balconies can be analyzed. The plane sweep operates with multiple planes that lie parallel to a key-plane. A *key-plane* is the approximate facade-plane. Additional planes are set parallel to the key-plane about one pixel apart (in our test area, this is at 10 cm) in both directions from the key-plane (see figure 4).

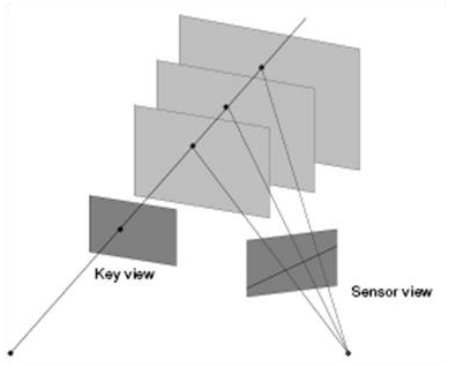


Figure 4: Plane sweeping principle. The homography between the facade's reference plane and the sensor view varies for different depths. (Zach, 2007).

If the plane at a certain depth passes exactly through parts of the object's surface to be reconstructed, a match will exist between the relevant parts of the new sensor view and the key view, the match being computed as a correlation. The sensor images are warped onto the current 3D key plane using the projective transformation.

After projecting a sensor image onto the current plane hypothesis, a correlation score for the current sensor view is calculated. The final correlation score of the current plane hypothesis is achieved by integrating all overlapping sensor views. For the accumulation of the single image correlation scores a simple additive blending operation is used. We repeat this process for all parallel planes and all corresponding images. The results of this calculation are n_k matching probabilities for every pixel x(i,j) of a facade for all n facade planes. Figure 5 illustrates the result of this correlation for 4 different planes.

In a next step we determine the depth map of a facade using a total generalized variation TGV multi labelling approach proposed by Pock et al. (2008). Figure 6 shows the resulting raw 3D point cloud and the 3D point cloud overlaid with RGB photo texture.

Figure 5: Correlation coefficients calculated for 4 different planes visualized as binary images (white areas have the largest correlation values).

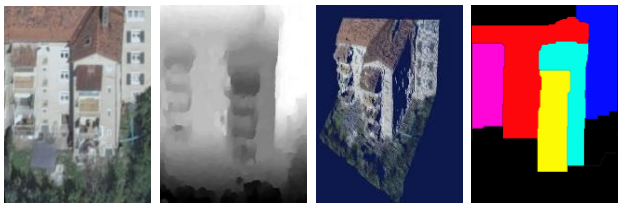


Figure 6: Reconstructed building facades using plane sweeping. (a) key view of facade (b) raw 3D point cloud (c) 3D point cloud overlaid with RGB information (d) segmentation of that point cloud into facade areas belonging to specific vertical planes.

The method produces a 3D point cloud that can now be used to determine if a facade is planar or complex depending on how many points of the 3D point cloud lie within a certain range of a regression plane. One now has to analyze the 3D points with the goal of segmenting the facade into its planar sub-facades. The problem is one of detecting planes in 3D point clouds for which various solutions exist.

The 3d point cloud is projected into the horizontal xy-plane and will present a footprint of the facade. We thereby reduce the point cloud from 3D to 2.5D, because we just use the most common depth values for every facade.

The projection is by column and starts by eliminating all outliers for every column and by searching for local maxima in each individual column. Then these values are weighted according to their appearance in the column. In a next step we look for maxima in the xy- plane (footprint of facade, see figure 7). These maxima in the xy-plane are the major facade planes of one facade. These resulting multiple planes do represent sub-facades ready for the detection of floors and windows.

The xy-plane contains the footprint. It is the basis for dealing with facade details and masonry of a building. We want to detect balconies. This is possible using the lower weights of the areas that are probably balconies, as shown in figure 7a (green highlighted profiles). Figure 8 illustrates a complex building facade, the footprint of the building and the effect of the elimination of the balconies and roof overhangs.

Figure 7: Plane detection for a complex facade; (a) key-view of a facade and projection in xy-plane (red lines are strongly weighted (>66% of points lie in this plane); (b) determined major facade planes (segmented areas with balconies are marked in red)

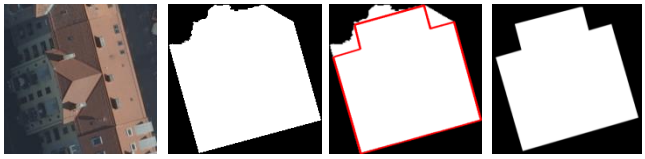


Figure 8: (a) building visible in vertical aerial image (b) segmentation result (c) in red modified building outline using 3D facade reconstruction (d) enhanced building footprint.

3.2 Experimental Results

For the evaluation of the 3D facades we have performed experiments in a test area of the city of Graz with a dimension of 400m x 400m with a Ground Sampling Distance of 10 cm and image overlaps in the range of 80% forward and 60% sideward. We have randomly selected 131 buildings with a total of 230 facades and have performed our plane detection method. Table 1 summarizes the results of this calculation depending on the number of aerial images used for the reconstruction. We achieve a detection rate of 70% for complex facades and a detection rate of 80% for planar facades.

	1 image	2 images		3 images		4 images		5 images		total	
images	5	68		77		52		28		230	
complex	0 %	64%	18/28	63%	15/24	78%	18/23	82%	9/11	**70%**	61/87
planar	0 %	75%	30/40	74%	39/53	86%	25/29	94%	16/17	**80%**	114/143
total	0 %	71%	48/68	70%	54/77	83%	43/52	90%	25/28	**76%**	175/230

Table 1: Evaluation of the facade plane detection using the column-wise approach

The main reasons why the plane detection fails are the combination of aerial images from two different flight strips for the computation and an inability to deal with dissimilarities in images taken with vastly different viewing perspectives. This is the case in about 20 facades of our data set. The results are noisy matching results that influence the outcome of the depthmap. One solution would be the reliance on imagery from just a single flight line. This would imply a very dense arrangement of flight lines and thus an increase of flying costs. At issue therefore would be further research and innovation in dealing with dissimilar facade images to obtain point clouds and matches.

By using these sub-facades for floor and window detection we achieve a detection rate for floors of 87% and for windows of 80%. Without the 3D reconstruction none of the complex facades could be interpreted correctly.

3.3 Discussion

We show that facades are being imaged usefully in vertical aerial photography. We also show that facades cannot be modelled as planes since there often is significant 3D structure. Use of the 3rd dimension for the interpretation of building facades is feasible with aerial photography and strongly improves the results. We achieve success rates of 87% for floor detection and 80% for window detection for facades that fail completely when a plane is assumed to be applicable.

We also show that it is possible to determine the extruding masonry of a building by eliminating balconies and roof overhangs. There are several avenues for improvements of the detail extraction from complex facades, and improvements of our understanding how well this works. First are data experiments in the form of a study with vertical aerial images with different GSD and different overlaps. Second is the ability of automatically recognizing occlusions and then responding to occluded facades by taking advantage of the overlapping images and their multitude of viewing angles. Innovations in window recognition will then become relevant, as will be site-dependent approaches to architectural styles.

4. CHARACTERIZATION OF BUILDING ROOFS USING AERIAL IMAGES

4.1 General Approach

Roofscapes offer similar complexities as facades, with predominant planes and multiple structures extruding from, sometimes also intruding into these planes. Nowadays, the preferred data source is LiDAR, and most of the recent literature on roof analysis is LiDAR-related. However, digital aerial photography is available and can be used. At issue is the segmentation of extended roofscapes into individual roofs, the measurement of the predominant roof planes and then the mapping of the 3D details. We build an approach within an overall framework introduced by Meixner et al. (2010). It employs vertical aerial images in order to characterize real properties.

After pre-processing the data to segment them into point clouds per individual property, and classification of roof areas, we move on to the characterization of the single building roofs. The interpretation of building roofs consists of 3 major processing steps:

- DSM Smoothing
- Plane Detection
- Roof Segmentation

4.2 DSM Smoothing

Photogrammetrically measured elevation data (range data) are noisy at the pixel to sub-pixel level and therefore may not easily be interpreted. For this reason and to accelerate the plane detection we smooth the elevation or range data using total generalized variation TGV using the approach developed by Pock et al. (2011). The result is illustrated in figure 9.

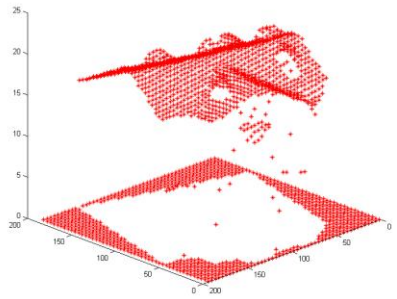


Figure 9: Smoothed point cloud of building roof (GSD 10cm); Dimensions: 200 x 200 pixels

4.3 Plane Detection

Roof planes are found using the "J-Linkage" method introduced by Toldo et al. (2008) that resembles the RANSAC method. It starts by random sampling where model hypotheses are generated. The essential difference to RANSAC is that minimal sets are constructed in a way that neighboring points are selected with higher probability. RANSAC treats all points the same. After all hypotheses are created, a preference set (set of hypothesis it prefers) is created for each point. Points that belong to the same structure have a similar preference set, meaning they are close in the conceptual space. To find the models "J-Linkage" uses an agglomerative clustering procedure, where at each step the two clusters with the minimum pairwise distance are merged. This distance reaches from 0 (identical sets) to 1 and just elements are linked together whose preference sets overlap. Figure 10 illustrates the result of this plane detection.

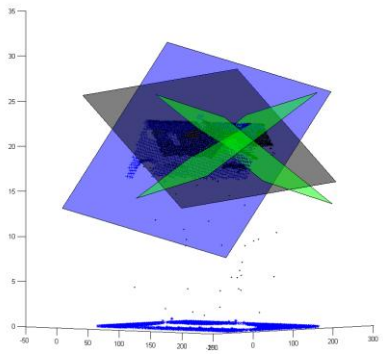


Figure 10: Point cloud from figure 9 highlighted in blue; the two major roof planes are marked in blue and black and the two smaller planes in green.

4.4 Roof Segmentation

After we have detected all roof planes we have to label them. Therefore we segment the roofs into 3 different classes, namely major roof planes, minor roof planes and superstructures. Using the major roof planes we obtain information about the style of the roof.

We determine if a plane segment is linked to any other segment to build a "region". Then the size of the resulting region or assembly of plane segments gets considered. Depending on the size of the regions with respect to the overall size of the roof we assign each region to an appropriate category. A refinement step serves to look at smaller plane segments associated perhaps with superstructures. Depending on their height values with respect to the neighbouring pixels these smaller regions are classified as part of a smaller plane (regions without height continuities at the borders), superstructures or are eliminated entirely. To achieve meaningful results we differentiate of course between height discontinuities at the edges of the roof and within the roof by using the information about a building from the building classification.

Of particular interest in our case are superstructures because they give us information about the use of the roof and building. We divide the superstructures into three groups: dormer windows, chimneys and other structures. For the example of chimneys all 4 edges have height discontinuities and that the maximum height is not lower than the height of the roof's ridge. By contrast the dormer windows have height discontinuities on at last three edges. Moreover their area is much larger than the area of the chimneys and the geometric form is more "quadratic" than elongated. Chimneys have usually smaller and narrower forms (< 0.5m width). Figure 11 illustrates the segmented roofscapes for two buildings.

Figure 10: Extract of an orthophoto of our test data set and overlaid Segmentation of roofscape into different roof categories (red: major planes; yellow: minor planes; green: chimneys; blue: dormer windows). [Source: Detail of the Graz test dataset]

4.5 Experiments

The Graz test site has 186 different buildings. A random selection of 20 buildings from this dataset serves as the basis for a segmentation experiment.

	Major planes	Smaller planes	Dormer windows	Chimneys	Other structures
Total planes	43	12	32	84	3
Detected planes	42	9	28	68	3
Detection rate [%]	98	75	88	80	100

Table 2: Detection rate for segmented roof structures

Major planes are being detected with a 98% success, 75% success was scored with smaller planes, 88% of the planes for the dormer windows and 80% of those for the chimneys were

correctly found. The overall success rate over all planes was at 87%. Misclassifications occur in complex roof structures like roof terraces or non-planar roof structures, thus in curved surfaces.

4.6 Discussion

We have presented in this section a method to segment building roofs into different roof categories: major plane, smaller planes, superstructures, and to interpret these in association with types of roofs, dormer windows, chimneys and other structures. We show that we can segment a roof with an accuracy of 87%. The method builds on point clouds and classifications of buildings from overlapping aerial photography with a GSD at 10 cm, and producing thus a point density of 25 pixels/m^2. This data source differs from the current predominant sensor for point clouds in the form of the airborne LiDAR.

Current results do encourage continued development of roof analysis work based on aerial photography. However, the experiments have shown that the proposed method should be improved. Particular difficulties occur when rather large dormers or roof gardens exist. Additionally, we need to extend the experimental effort to include different architectural styles and building uses to include coastal resort environments, historical small towns, alpine terrains, urban cores with skyscrapers and industrial zones.

5. CONCLUSION

GeoVirtual Environments include the advent of 3D virtual cities in the form of 3D models of the urban human habitat. The Internet as an initial inspiration to quickly develop 3D city models has put the focus in the visual appeal of the result, not on the ability to use the building data for any analysis.

We argue that this needs to change and that images need to serve to characterize the real properties, that building details be part of the data base and can be searched. Buildings should be found based not only on an address, but also on the number of floors, the size of facades, the number of windows and the architectural style of a building.

Aerial photography is a work horse for urban mapping and exists for all urban spaces. It contains information about facades and roofs that needs to get extracted. Initial work succeeds in finding relevant information with accuracies in the range of 88% and more. Aerial and street side imagery needs to be used cooperatively to overcome the limitations of each data source, such as occlusions in street side data or poor facade texture in aerial data.

LiDAR has not been addressed in this contribution, although LiDAR is a contender in any urban geo-sensing and is in everybody's mind when 3D point clouds are at issue. We did show that imagery is a valid and useful source of geometric 3D information of building facades and roofs, and by this we want to highlight that digital high overlap imagery produces point clouds for successful roof as well as facade analysis.

References

Hammoudi, K., Dornaika, F., Soheilian, B., Paparoditis, N.. A Pipeline for Modeling Urban Street Facades from Terrestrial Laser and Image Data. *in ICVSS International Computer Vision Summer School*, Sicily, Italy, to appear.

Meixner, P., Leberl, F. 2011. 3-Dimensional Building Details from Aerial Photography for Internet Maps. *Remote Sening.* 3, no. 4, pp. 721-751.

Pock, T., Schoenemann, T., Graber, G., Bischof, H., Cremers, D. 2008. A Convex Formulation of Continuous Multi-Label Problems. In: *Proceedings of the European Conference on Computer Vision* ECCV, pp. 792-805

Pock, T., Zebedin, L., Bischof, H. 2008 TGV - Fusion. In: *Rainbow of Computer Science*. Springer-Verlag. To appear.

Recky, M., Wendel, A., Leberl, F. 2011. Facade Segmentation in Multi-View Scenario. In: *Proceedings International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission* [3DIMPVT 2011], Hangzhou, PR China. Published by IEEE Computer Society, pp. 358-365.

Recky, M., Leberl, F. 2009. Semantic Segmentation of Street-Side Images. In: *Proceedings of the Annual OAGM Workshop*. Austrian Computer Society in OCG, pp. 271–282.

Toldo, R., Fusiello, A. 2008. Robust Multiple Structures Estimation with J-linkage. In: *Proceedings of the European Conference on Computer Vision* ECCV, Part 1, 537-547, ISBN: 978-3-540-88681-5

Wendel, A, Donoser, M., Bischof, H. 2010. Unsupervised facade segmentation using repetitive patterns. In: *Proceedings of the 32nd Annual Symposium of the German Association for Pattern Recognition* (DAGM'10), Springer LNCS 6376, pp. 51-60.

Wendel A. 2009. *Facade Segmentation from Streetside Image.* Master's Thesis, Graz University of Technology, 72 p.

Zach, C. (2007) *High Performance Modelling from Multiple Views using Graphics Hardware*. Doctoral Thesis, Graz University of Technology, 153 p.

IMPROVED BUILDING DETECTION USING TEXTURE INFORMATION

Mohammad Awrangjeb, Chunsun Zhang and Clive S. Fraser

Cooperative Research Centre for Spatial Information,University of Melbourne
Level 5, 204 Lygon Street, Carlton Vic 3053, Australia
Phone: +61 3 8344 9182, Fax: +61 3 9349 5185
Email: {mawr, chunsunz, c.fraser}@unimelb.edu.au

Working Groups III/4

KEY WORDS: Building, Detection, LIDAR, Orthoimage, Fusion, Texture, Classification.

ABSTRACT:

The performance of automatic building detection techniques can be significantly impeded due to the presence of same-height objects, for example, trees. Consequently, if a building detection technique cannot distinguish between trees and buildings, both its false positive and false negative rates rise significantly. This paper presents an improved automatic building detection technique that achieves more effective separation of buildings from trees. In addition to using traditional cues such as height, width and colour, the proposed improved detector uses texture information from both LIDAR and orthoimagery. Firstly, image entropy and colour information are jointly applied to remove easily distinguishable trees. Secondly, a voting procedure based on the neighbourhood information from both the image and LIDAR data is employed for further exclusion of trees. Finally, a rule-based procedure using the edge orientation histogram from the image is followed to eliminate false positive candidates. The improved detector has been tested on a number of scenes from three different test areas and it is shown that the algorithm performs well in complex scenes.

1 INTRODUCTION

Building detection from remotely sensed data has a number of practical applications including city planning, homeland security and disaster management. Consequently, a large number of building detection techniques have been reported over the last few decades. Since photogrammetric imagery and LIDAR (LIght Detection And Ranging) data have their own merits and demerits, the recent trend is to integrate data from both of these sources as a means of advancing building detection by compensating the disadvantages of one with the advantages of the other.

The success of automatic building detection is still largely impeded by scene complexity, incomplete cue extraction and sensor dependency of data (Sohn and Dowman, 2007). Vegetation, and especially trees, can be the prime cause of scene complexity and incomplete cue extraction. Image quality may vary for the same scene even if images are captured by the same sensor, but at different times. The situation also becomes complex in hilly and densely vegetated areas where only a few buildings are present, these being surrounded by trees. Important building cues can be completely or partially missed due to occlusions and shadowing from trees. Therefore, many existing building detection techniques that depend largely on colour information exhibit poor detection performance.

Application of a recently developed building detection algorithm (Awrangjeb et al., 2010a) has shown it to be capable of detecting buildings in cases where cues are only partially extracted. For example, if a section of the side of a roof (at least 3m long) is correctly detected, the algorithm can also detect all or part of the entire building. However, this detector does not necessarily work well in complex scenes when buildings are surrounded by dense vegetation and when they have the same colour as trees, or where trees are other than green.

This paper presents an improved detection algorithm that uses both LIDAR and imagery. In addition to exploiting height, width and colour information, it uses different texture information in order to differentiate between buildings and trees. Firstly, image entropy and colour information are employed together to remove the trees that are easily distinguishable. Secondly, a voting procedure that considers neighbourhood information is proposed for the further exclusion of trees. Finally, false positive detections are eliminated using a rule-based procedure based on the edge orientation histogram. The improved detector has been tested on a number of scenes covering three different test areas [1].

2 CUES TO DISTINGUISH TREES AND BUILDINGS

Cues employed to help distinguish trees from buildings include the following:

- **Height**: A height threshold (2.5m above ground level) is often used to remove low vegetation and other objects of limited height, such as cars and street furniture (Awrangjeb et al., 2010a). The height difference between first and last pulse DSMs (digital surface models) have also been used (Khoshelham et al., 2008).
- **Width, area and shape**: If the width or area of a detected object is smaller than a threshold, then it is removed as a tree (Awrangjeb et al., 2010a). A number of shape attributes can be found in (Matikainen et al., 2007).
- **Surface**: A plane-fitting technique has been applied to non-ground LIDAR points to separate buildings and trees (Zhang et al., 2006), and a polymorphic feature extraction algorithm applied to the first derivatives of the DSM in order to estimate the surface roughness has also been employed (Rottensteiner et al., 2007).
- **Colours**: While a high NDVI (normalised difference vegetation index estimated using multispectral images) value represents a vegetation pixel, a low NDVI value indicates

[1] This paper is a condensed version of (Awrangjeb et al., 2011).

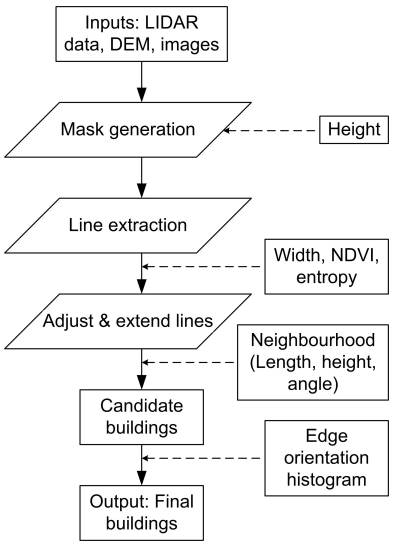


Figure 1: Flow diagram of the improved building detection technique.

a non-vegetation pixel. This cue, although frequently used, has been found unreliable even in normal scenes where trees and buildings have distinct colours (Awrangjeb et al., 2010a). K-means clustering was applied on multispectral images to obtain spectral indices for clusters like trees, water and buildings (Vu et al., 2009). Colour invariants have also been used (Shorter and Kasparis, 2009). A number of other cues generated from colour image and height data can be found in (Matikainen et al., 2007, Salah et al., 2009).

- **Texture**: When objects have similar spectral responses, the grey level co-occurrence matrix (GLCM) can be estimated from the image to quantify the co-occurrence probability (Chen et al., 2006). Some GLCM indices, eg mean, standard deviation, entropy and homogeneity, have been applied to both height and image data in order to classify buildings and trees (Salah et al., 2009, Matikainen et al., 2007).

- **Training pixels**: Training pixels of different colours from roofs, roads, water, grass, trees and soil have been used for classification (Lee et al., 2003).

- **Filtering**: Morphological opening filters have been employed to remove trees attached to buildings (Yong and Huayi, 2008).

- **Others**: Segmentation of LIDAR intensity data can also be used to distinguish between buildings and trees (Maas, 2001). The density of raw LIDAR data has also been employed (Demir et al., 2009).

3 IMPROVED BUILDING DETECTION

The proposed improved detector employs a combination of height, width, angle, colour and texture information with the aim of more comprehensively separating buildings from trees. Although cues other than texture were used in the earlier version of the detector, the improved formulation makes use of additional texture cues such as entropy and the edge orientation histogram at four stages of the process, as shown in Fig. 1. Different steps of the detection algorithm have been presented in (Awrangjeb et al., 2010a). This paper focuses on how texture, dimensional and colour information can be applied jointly in order to better distinguish between buildings from trees. The setup of different threshold values are discussed in (Awrangjeb et al., 2011).

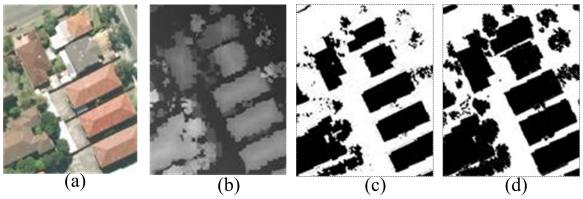


Figure 2: (a) Image of a test scene, (b) corresponding LIDAR data (in gray-scale), (c) primary mask and (d) secondary mask.

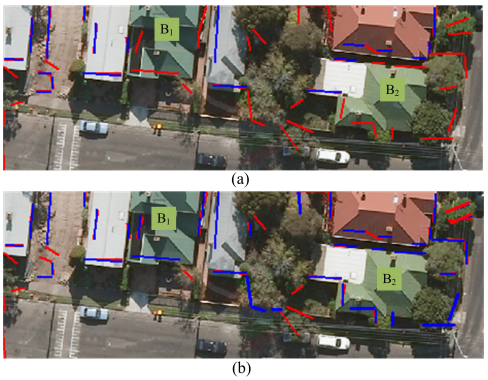


Figure 3: Detection of green buildings: (a) the NDVI information alone missed green buildings whereas (b) combined NDVI and entropy information detects green buildings. 'Blue' lines are accepted, 'red' represents rejected.

3.1 Application of Height Threshold

A height threshold $T_h = H_g + 2.5$m, where H_g represents the ground height, is applied to the raw LIDAR data and two building masks are created – the primary M_p and secondary M_s masks (Awrangjeb et al., 2010a). This threshold removes low height objects (grounds, grass, roads, cars etc.) and preserves non-ground points (trees and buildings). The corresponding DEM height for a given LIDAR point is used as the ground height. If there is no corresponding DEM height for a given LIDAR point, the average DEM height in the neighbourhood is used. Fig. 2 shows the two extracted masks for a scene.

3.2 Use of Width, NDVI and Entropy

The black areas in M_p are either buildings, trees or other elevated objects. Line segments around these black shapes in M_p are formed, and in order to avoid detected tree-edges, extracted lines shorter than the minimum building width $L_{min} = 3$m are removed. Trees having small horizontal area are thus removed.

The mean of the NDVI value is then applied, as described in (Awrangjeb et al., 2010a), to eliminate trees having large horizontal area. However, the NDVI has been found to be an unreliable cue even in normal scenes where trees and buildings have distinct colours (Rottensteiner et al., 2007, Awrangjeb et al., 2010a). In addition, it cannot differentiate between trees and green buildings. Fig. 3(a) shows an example where a green building B_1 cannot be detected at all since all lines around it are rejected. However, green building B_2 can be partially detected because it has a white coloured roof section. In some areas there may be non-green buildings having the same colour as trees, especially when leaves change colour in different seasons. In such cases, the removal of trees based on the NDVI will result in many buildings also being removed. Detection of these same buildings will likely also lead to detection of trees.

If the mean NDVI is above the NDVI threshold at any side of a line segment, a further test is performed before removing this

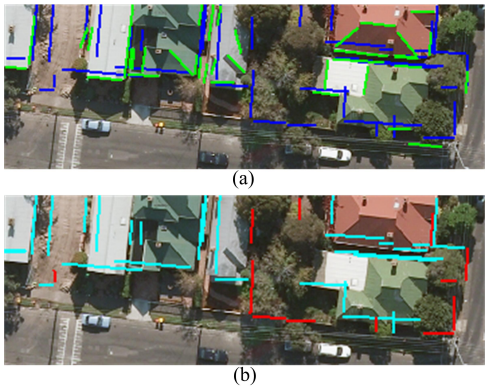


Figure 4: Use of neighbourhood information to remove tree-edges: (a) before voting: 'blue' represents lines from the primary mask after the extending procedure and 'green' represents lines from the image and (b) after voting: 'cyan' represents accepted lines after the voting procedure and 'red' represents rejected lines.

line segment as a tree-edge. This test checks whether the average entropy is more than the entropy threshold $T_{ent} = 30\%$. If the test holds, the line segment is removed as a tree edge, otherwise it is selected as a building edge. Fig. 3(b) shows that the green buildings B_1 and B_2 can be fully detected using this approach. In addition, some of the trees subject to shadowing and self-occlusion are also detected.

3.3 Voting on the Neighbourhood Information

The joint application of NDVI and entropy can remove some large trees; however, in the case when there are shadows and self-occlusions within trees, difficulties with the approach can be expected. Therefore, for each of the extended lines a voting procedure based on the information within the neighbourhood of that line is followed.

All the extracted and extended lines that reside around the same black shape in the primary mask M_p fall into the same neighbourhood. Let $\Omega = \{l_i\}$, $0 \leq i \leq n_t$ be such a neighbourhood obtained after the application of the width threshold L_{min} in the previous section, where l_i indicates an extracted line, its length $L_{l_i} \geq$ 3m, and there are a total of n_t extracted lines. Furthermore, let n_e lines, out of n_t extracted lines in Ω, survive after the extending procedure discussed above, with the average length of these being $L_{\Omega,avg}$. We also consider the longest image line, extracted from the grey-scale orthoimage, which resides around l_i. The longest local image line ℓ_i for l_i within a rectangular area of width 3m around l_i is obtained. Let the length of ℓ_i be L_{ℓ_i}. In some cases, no ℓ_i may be found due to poor image contrast or if l_i is a tree edge. Fig. 4(a) shows the extended lines from M_p and the accepted lines from the orthoimage.

For each line l_i in the proposed voting procedure, four votes v_k, $0.0 \leq v_k \leq 1.0$ are cast by exploiting its neighbourhood information as follows:

- $v_1 = 1.0$, if $L_{l_i} \geq L_{\Omega,avg}$; else $v_1 = \frac{L_{\Omega,avg} - L_{l_i}}{L_{\Omega,avg}}$.
- $v_2 = \frac{\Theta - \theta_i}{\Theta}$, where θ_i is the adjustment angle between l_i and the longest line in Ω, which was used as the base line in the adjustment procedure, and $\Theta = \frac{\pi}{8}$ is the angle threshold used in the adjustment procedure (Awrangjeb et al., 2010a).
- $v_3 = \frac{n_e}{n_t}$. This is based on the observation that line segments around a building are more likely to be adjusted, which means that they are either parallel or perpendicular to the base line around the same black shape in M_p.
- $v_4 = 1.0$ if $L_{\ell_i} \geq 2L_{min}$; else $v_4 = \frac{L_{\ell_i} - L_{min}}{L_{min}}$. If there is no image line found around l_i, then $v_4 = 0.0$.

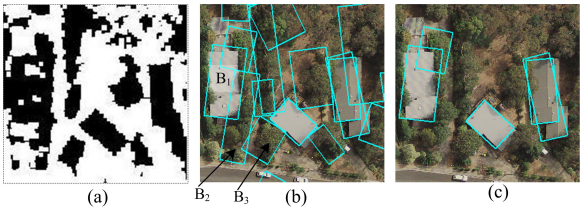


Figure 5: A complex scene: (a) primary mask, (b) detected candidate buildings with a large number of false detections and (c) detected final buildings after removing false positives.

The voting procedure is executed for n_e lines in Ω. A line l_i is designated a building edge if it obtains a majority vote. This means that the mean of v_k, $1 \leq k \leq 4$, is greater than 0.50. Fig. 4(b) shows that the majority of tree edges can be removed by applying the voting procedure. A candidate building set is then obtained using the extended lines that survive the voting procedure (Awrangjeb et al., 2010a).

In areas with dense vegetation, the black shapes of buildings and nearby trees are not separable and consequently a building may be connected with another building a few metres away (see Fig. 5). If the connected buildings are not parallel to each other, then the improved adjustment procedure will likely still fail. This is why in the improved detection algorithm, the adjustment and voting procedure is available as an optional step, the choice of which will depend upon vegetation density. In either case, there may be some false buildings present in the candidate building set, as shown in Fig. 5(b). A procedure utilising the edge orientation histogram from the orthoimage is then applied in order to remove false positives.

3.4 Application of Edge Orientation Histogram

Following the detection of candidate buildings, a gradient histogram is formed using the edge points within each candidate building rectangle. Edges are first extracted from the orthophoto using an edge detector and short edges (less than 3m in length) are removed. Each edge is then smoothed and the gradient (tangent angle) is calculated on each point using the first order derivatives. The gradient will be in the range $[-90°, +90°]$. A histogram with a successive bin distance of $D_{bin} = 5°$ is formed using the gradient values of all edge points lying inside the candidate rectangle.

Rectangles containing the whole or major part of a building should have one or more significant peaks in the histogram, since edges detected on building roofs are formed from straight line segments. All points on an apparent straight line segment will have a similar gradient value and hence will be assigned to the same histogram bin, resulting in a significant peak. A significant peak means the corresponding bin height is well above the mean bin height of the histogram. Since edge points whose gradient falls into the first (at $-90°$ to $-85°$) and last (at $85°$ to $90°$) bins have almost the same orientation, located peaks in these two bins are added to form a single peak.

Fig. 6 illustrates three gradient histogram functions and mean heights for candidate buildings B_1, B_2 and B_3 in Fig. 5(b). Two bins at $\pm 90°$ basically form one bin, because lines in these two

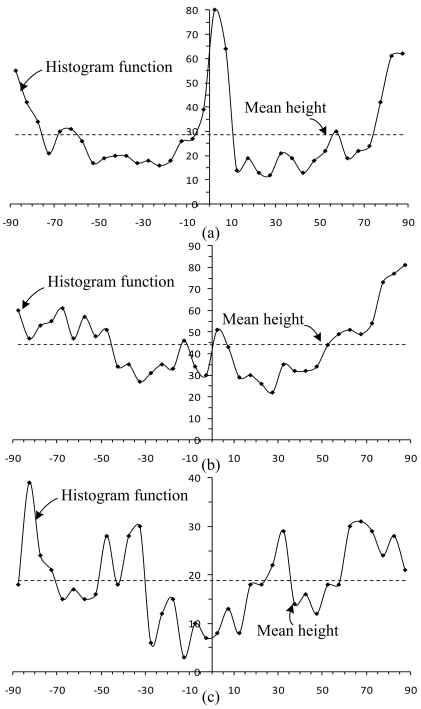


Figure 6: Gradient histogram functions and means for rectangles (a) B_1, (b) B_2 and (c) B_3 in Fig. 5(b): x-axis is in degrees and y-axis is in pixels (bin heights).

bins are perpendicular to the *x*-axis and reside above & below this axis. Therefore, these can be a peak at either of these bins and their heights can be accumulated to form a single peak. Fig. 6(a) shows that B_1 has two significant peaks: 80 pixels at 0° and 117 $(55 + 62)$ pixels at $\pm 90^\circ$, these being well above the mean height of 28.6 pixels. The two significant peaks separated by 90° strongly suggest that this is a building. From Fig. 6(b) it can be seen that B_2 has one significant peak at $\pm 90^\circ$ but a number of insignificant peaks. This points to B_2 being partly building but mostly vegetation, which is also supported by the high mean height value. With the absence of any significant peak, but a number of insignificant peaks close to the mean height, Fig. 6(c) indicates that B_3 is comprised of vegetation. Although there may be some significant peaks in heavily vegetated areas, a high average height of bins between two significant peaks can be expected. Note that the orthophoto resolution in this case was 10cm, so a bin height of 80 pixels indicates a total length of 8m from the contributing edges.

The observations above support the theoretical inferences. In practice, however, detected vegetation clusters can show the edge characteristics of a building, and a small building having a flat roof may not have enough edges to show the required peak properties. As a result, some true buildings can be missed, while some false buildings may be detected. A number of precautions can be formulated in order to minimize the occurrence of false detections.

Two types of histograms are formed using edges within each detected rectangle. In the first type, one histogram considers all the edges collectively, and in the second type histograms for individual edges whose length is at least L_{min} are formed. Let the collective histogram be symbolized as H_{col}, with an individual histogram being indicated by H_{ind}. Tests on H_{col} and H_{ind} can be carried out to identify true buildings and remove trees. If a detected rectangle passes at least one of the following tests it is selected as a building, otherwise it is removed as vegetation.

1. *Test 1*: H_{col} has at least two peaks with heights of at least $3L_{min}$ and the average height of bins between those peaks is less than $2L_{min}$. This test ensures the selection of a large building, where at least two of its long perpendicular sides are detected. It also removes vegetation where the average height of bins between peaks is high.

2. *Test 2*: The highest bin in H_{col} is at least $3L_{min}$ in height and the aggregated height of all bins in H_{col} is at most 90m. This test ensures the selection of a large building where at least one of its long sides is detected. It also removes vegetation where the aggregated height of all bins is high.

3. *Test 3*: H_{col} has at least two peaks with heights of at least $2L_{min}$, and the highest bin to mean height ratio R_{Mm1} is at least 3. This test ensures the selection of a medium size building, where at least two of its perpendicular sides are detected. It also removes vegetation where the highest bin to mean height ratio is low.

4. *Test 4*: The highest bin in H_{col} has a height of at least L_{min} and the highest bin to mean height ratio R_{Mm2} is at least 4. This test ensures the selection of a small or medium size building where at least one of its sides is at least partially detected. It also removes small to moderate sized vegetation areas where the highest bin to mean height ratio is low.

5. *Test 5*: The highest bin in H_{ind} has a height of at least L_{min} and the aggregated height of all bins in H_{col} is at most 90m. This test ensures the selection of buildings which are occluded on at most three sides.

6. *Test 6*: The ratio R_{aTp} of the detected rectangular area to the number of texture pixels (N_{Tp}, the aggregated height of all bins in H_{col}) is at least 45. This test ensures the selection of all buildings which are at least partially detected but the roof sides are missed.

The application of these tests on the complex scene in Fig. 5(b) produces the result shown in Fig. 5(c). Note that for simple scenes with small amounts of vegetation, the NDVI and entropy together can successfully remove most trees so subsequent application of the voting procedure and edge orientation histogram can be considered as optional, leading to a saving of computation time.

4 RESULTS AND DISCUSSIONS

The threshold-free evaluation system involved in the performance study conducted makes one-to-one correspondences using nearest centre distances between detected and reference buildings. The descriptor 'threshold-free' means the evaluation system does not involve any thresholds based on human choice. Some 15 evaluation indices in three categories, namely object-based, pixel-based and geometric, have been employed. Whereas pixel-based evaluation considers only spectral properties in the imagery, object-based evaluation takes into account spatial and contextual properties in both the imagery and LIDAR data. The root mean square positional discrepancy value (RMSE) is employed to quantify the geometric accuracy. The detailed procedure of the threshold-free evaluation system and the evaluation indices can be found in (Awrangjeb et al., 2010b).

The test data sets employed cover three suburban areas in Australia, Fairfield, NSW; Moonee Ponds, Victoria and Knox, Victoria. The Fairfield data set covers an area of 588m $\times$ 417m and contains 370 buildings, Moonee Ponds covers 447m $\times$ 447m and

has 250 buildings and Knox covers 400m × 400m and contains 130 buildings. Fairfield contains many large industrial buildings and in Mooney Ponds there were some green buildings. Knox can be characterized as outer suburban with lower housing density and extensive tree coverage that partially covers buildings. In terms of topography, Fairfield and Mooney Ponds are relatively flat while Knox is quite hilly.

LIDAR coverage comprised last-pulse returns with a point spacing of 0.5m for Fairfield, and first-pulse returns with a point spacing of 1m for Moonee Ponds and Knox. For Fairfield and Knox, RGB colour orthoimagery was available, with resolutions of 0.15m and 0.1m, respectively. Moonee Ponds image data comprised RGBI colour orthoimagery with a resolution of 0.1m. Bare-earth DEMs of 1m horizontal resolution covered all three areas.

Reference data sets were created by monoscopic image measurement using the Barista software [2]. All rectangular structures, recognizable as buildings and above the height threshold T_h, were digitized. The reference data included garden sheds, garages, etc. These were sometimes as small as 10m^2 in area.

Tables 1 to 3 show results of the object-based, pixel-based and geometric accuracy evaluations of the improved building detection algorithm in the three test areas. A visual illustration of sample building detection results are shown in Fig. 7. The improved algorithm produced moderately better performance than the original in all three evaluation categories within both Fairfield and Mooney Ponds. The better performance was mainly due to proper detection of large industrial buildings in Fairfield, detection of some green buildings in Mooney Ponds, and elimination of trees in both Fairfield and Mooney Ponds.

In Knox, the improved algorithm exhibited significantly better performance over the original, due to two main reasons. Firstly, the improved algorithm better accommodated the dense tree cover and randomly oriented buildings that characterized the Knox data. Fairfield and Mooney Ponds on the other hand are low in vegetation cover and buildings are generally well separated and more or less parallel or perpendicular to each other. Secondly, the improved algorithm showed its merits in better handling varying topography. Knox is a hilly area (maximum height $H_M = 270$m and minimum height $H_m = 110$m), whereas Fairfield ($H_M = 23$m and $H_m = 1$m) and Mooney Ponds ($H_M = 43$m and $H_m = 23$m) are moderately flat.

The original algorithm detected a large number of false buildings in Knox, as illustrated in Figs. 7 (a) and (c). Moreover, many buildings detected with the original algorithm were not properly aligned. Consequently, in object-based evaluation, 56% quality was observed with 77% completeness and 67% correctness. The reference cross-lap rate was above 85%, with 39% detection overlap rate. In pixel-based evaluation, 27% quality was found with 44% completeness and 42% correctness. The area omission error was more than 50% and both branching and miss factors were above 120%. The geometric accuracy was no better than 33 pixels.

In contrast, as shown for Knox in Figs. 7 (b) and (d), the improved detector removed a large number of false buildings using its orientation histogram. In object-based evaluation, when compared to the original algorithm, the quality increased to 82%, a 26% rise. The detection overlap rate decreased to 13% and the reference cross-lap rate reduced to 62%. In pixel-based evaluation, again when compared to the original algorithm, the quality went up to 39%, a 12% growth, while the branching factor declined dramatically to 52% and the miss factor was also moderately improved to 104%. Geometric accuracy improved to 20 pixels, or by approximately 50%.

[2]The Barista Software, www.baristasoftware.com.au, May 2011.

Table 1: Object-based evaluation results in percentages (C_m = completeness, C_r = correctness, Q_l = quality, M_d = multiple detection rate, D_o = Detection overlap rate, C_{rd} = detection cross-lap rate and C_{rr} = reference cross-lap rate).

Scenes	C_m	C_r	Q_l	M_d	D_o	C_{rd}	C_{rr}
Fairfield	95.1	95.4	92.2	2.7	8.6	3.5	9.7
MPonds	94.5	95.3	89.2	6.2	13.1	7.3	17.5
Knox	93.2	87.2	82.0	9.3	12.8	23.3	61.6
Average	**94.0**	**91.3**	**86.4**	**6.9**	**11.9**	**14.4**	**37.6**

Table 2: Pixel-based evaluation results in percentages (C_{mp} = completeness, C_{rp} = correctness, Q_{lp} = quality, A_{oe} = area omission error, A_{ce} = area commission error, B_f = branching factor and M_f = miss factor).

Scenes	C_{mp}	C_{rp}	Q_{lp}	A_{oe}	A_{ce}	B_f	M_f
Fairfield	83.2	84.5	72.4	15.3	12.5	13.5	20.3
MPonds	87.2	85.4	75.3	12.7	13.2	16.7	17.3
Knox	49.0	65.9	39.1	51.0	30.9	51.8	104.0
Average	**73.1**	**78.6**	**62.3**	**26.3**	**18.9**	**27.3**	**47.2**

In object-based evaluation, the improved algorithm offered on average across the three data sets a more than 10% increase in completeness and correctness and a 15% increase in quality. Multiple detection and detection overlap rates were also low. In pixel-based evaluation, there was also a reasonable rise in completeness (4%), correctness (10%) and quality (7%). Area omission and commission errors were less than those obtained with the original algorithm. In addition, there was a 5 pixel improvement in geometric accuracy.

5 CONCLUSIONS

This paper has presented an improved automatic building detection technique that exhibits better performance in separating buildings from trees. In addition to employing height and width thresholds and colour information, it uses texture information from both LIDAR and colour orthoimagery. The joint application of measures of entropy and NDVI helps in the removal of vegetation by making trees more easily distinguishable. The voting procedure incorporates neighbourhood information from the image and LIDAR data for further removal of trees. Finally, a rule-based procedure based on the edge orientation histogram from the image edges assists in eliminating false positive building candidates. The experimental results reported showed that while the improved algorithm offered moderately enhanced performance in Fairfield and Mooney Ponds, it yielded a very significant improvement in performance in Knox across all three evaluation categories.

ACKNOWLEDGEMENTS

The authors would like to thank the Department of Sustainability and Environment (www.dse.vic.gov.au) for providing the LIDAR data and orthoimagery for the Mooney Ponds and Knox test sites.

REFERENCES

Awrangjeb, M., Ravanbakhsh, M. and Fraser, C. S., 2010a. Automatic detection of residential buildings using lidar data and multispectral imagery. ISPRS Journal of Photogrammetry and Remote Sensing 65(5), pp. 457–467.

Figure 7: Building detection by the previous (left) and the improved (right) algorithms on two samples from Knox.

Table 3: Geometric accuracy.

Scenes	metres	pixels
Fairfield	2.4	16.0
Mooney Ponds	1.6	16.0
Knox	2.0	20.0
Average	**2.0**	**17.3**

Awrangjeb, M., Ravanbakhsh, M. and Fraser, C. S., 2010b. Building detection from multispectral imagery and lidar data employing a threshold-free evaluation system. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 38(part 3A), pp. 49–55.

Awrangjeb, M., Zhang, C. and Fraser, C. S., 2011. Infomration removed to facilitate the blind review system. Submitted to ISPRS Journal of Photogrammetry and Remote Sensing.

Chen, L., Teo, T., Hsieh, C. and Rau, J., 2006. Reconstruction of building models with curvilinear boundaries from laser scanner and aerial imagery. Lecture Notes in Computer Science 4319, pp. 24–33.

Demir, N., Poli, D. and Baltsavias, E., 2009. Extraction of buildings using images & lidar data and a combination of various methods. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 38(part 3/W4), pp. 71–76.

Khoshelham, K., Nedkov, S. and Nardinocchi, C., 2008. A comparison of bayesian and evidence-based fusion methods for automated building detection in aerial data. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 37(part B7), pp. 1183–1188.

Lee, D., Shan, J. and Bethel, J., 2003. Class-guided building extraction from ikonos imagery. Photogrammetric Engineering and Remote Sensing 69(2), pp. 143–150.

Maas, H. G., 2001. The suitability of airborne laser scanner data for automatic 3d object reconstruction. In: Proc. 3rd International Workshop on Automatic Extraction of Man-Made Objects from Aerial and Space Images, Ascona, Switzerland, pp. 345–356.

Matikainen, L., Kaartinen, H. and Hyypp, J., 2007. Classification tree based building detection from laser scanner and aerial image data. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 36(part 3/W52), pp. 280–287.

Rottensteiner, F., Trinder, J., Clode, S. and Kubik, K., 2007. Building detection by fusion of airborne laser scanner data and multi-spectral images : Performance evaluation and sensitivity analysis. ISPRS Journal of Photogrammetry and Remote Sensing 62(2), pp. 135–149.

Salah, M., Trinder, J. and Shaker, A., 2009. Evaluation of the self-organizing map classifier for building detection from lidar data and multispectral aerial images. Journal of Spatial Science 54(2), pp. 1–20.

Shorter, N. and Kasparis, T., 2009. Automatic vegetation identification and building detection from a single nadir aerial image. Remote Sensing 1(4), pp. 731–757.

Sohn, G. and Dowman, I., 2007. Data fusion of high-resolution satellite imagery and lidar data for automatic building extraction. ISPRS Journal of Photogrammetry and Remote Sensing 62(1), pp. 43–63.

Vu, T., Yamazaki, F. and Matsuoka, M., 2009. Multi-scale solution for building extraction from lidar and image data. International Journal of Applied Earth Observation and Geoinformation 11(4), pp. 281–289.

Yong, L. and Huayi, W., 2008. Adaptive building edge detection by combining lidar data and aerial images. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 37(part B1), pp. 197–202.

Zhang, K., Yan, J. and Chen, S. C., 2006. Automatic construction of building footprints from airborne lidar data. IEEE Trans. on Geoscience and Remote Sensing 44(9), pp. 2523–2533.

RANGE AND IMAGE DATA INTEGRATION FOR MAN-MADE OBJECT RECONSTRUCTION

F. Nex , F. Remondino

3D Optical Metrology, Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy
<franex, remondino>@fbk.eu, http://3dom.fbk.eu

Working Groups I/2, III/1, III/3, III/4

KEY WORDS: Integration, Image, Matching, LiDAR, Automation, Aerial, Close range

ABSTRACT:

The extraction of information from image and range data is one of the main research topics. In literature, several papers dealing with this topic has been already presented. In particular, several authors have suggested an integrated use of both range and image information in order to increase the reliability and the completeness of the results exploiting their complementary nature. In this paper, an integration between range and image data for the geometric reconstruction of man-made object is presented. In particular, the focus of this paper is on the edge extraction procedure performed in an integrated way exploiting the information provided by both range and image data. Both terrestrial and aerial applications have been analysed for the façade extraction in terrestrial acquisitions and the roof outline extraction from aerial data. The algorithm and the achieved results are described and discussed in detail.

1. INTRODUCTION

The extraction of precise and dense point clouds is nowadays possible using both LiDAR and image matching techniques. Nowadays, LiDAR and photogrammetric point clouds can indeed be considered comparable in most cases in terms of density and accuracy, as asserted in several papers both in terrestrial (Remondino et al., 2008; Hiep et al., 2009) and aerial applications (Paparoditis et al., 2006; Hirshmuller, 2008; Gehrke et al., 2010).

In typical mapping and modelling applications, once a point cloud (usually several millions of points) has been extracted, only the first (and shortest) part of the work has been completed. It is afterward required to process them in order to extract metric information (such as shapes, surface normal vectors, dimensions, polylines, etc.) of different objects (façades, buildings, streets, etc.) necessary to achieve the final product (3D model, drawing, thematic map, etc.). In some way, the classification, segmentation, modelling and in general the "understanding" of an unstructured point cloud in an almost automated way and without loss of accuracy is the real challenge to be faced nowadays by researchers. In the literature several papers deal with such topics. Some contributions considered as input data only images (Zebelin et al., 2006), others considered range information (Habib et al., 2009; Sampath and Shan, 2009; Pu and Vosselmann, 2009) and a growing number of papers rely on the integration of different data sources (Demir et al., 2009) and in particular from both range and image data (Alshawabkeh, 2006; Awrangjeb, et al. 2010; Habib et al., 2010). The single-technique approaches usually provide good results in very specific applications, while they are unable to be adapted to operative conditions far from their original use. On the other hand, the multi-technique solutions seem to be more versatile and able to achieve good results for a wider range of applications exploiting the complementary nature of range and image data. In particular, the integration of range and image data has shown promising results both in terrestrial and aerial applications for 3D modeling and mapping purposes (Alshawabkeh, 2006; Habib et al. 2010) with a great improvement for façade modelling.

Generally the main problem is the correct extraction and 3D reconstruction of continuous and reliable edges. A complete edge extraction allows indeed to reconstruct the geometry of the surveyed objects in a better way (Nex, 2010). In this paper, a different edge extraction approach is presented. Assuming to have, for a defined area of interest, a point cloud (from range sensors or image matching) and a set of images, an integrated methodology is applied in order to retrieve the main geometric discontinuities of the scene which are the base for successive drawing or mapping purposes. Range sensors generally suffer in measuring the scene edges. On the other hand, despite edges are clearly measurable in the images, automated matching algorithms are usually not tailored to extract edges but they are matched only as integration to other image primitives. Therefore the proposed method extracts a set of reliable and continuous edges combining the available point cloud and images. The main goal is to obtain only the edges strictly requested in the geometric reconstruction of the surveyed scene. A data driven approach is used in order to keep the maximum flexibility and detect discontinuities of generic shape. The final results show a more complete 3D reconstruction of man-made objects. The reported case studies are related to objects, in terrestrial (building façades) and aerial acquisitions (roof outlines).

In the following section a brief overview of the workflow is given with the integrated edge extraction step presented more in detail. Then, the achieved results in the 3D object reconstruction will be shown. Finally, conclusions and future developments will be discussed.

2. ALGORITHM'S OVERVIEW

The proposed algorithm processes image and range information in order to extract geometric primitives useful for a more

complete and detailed reconstruction of man-made objects like building façades or roofs. A point cloud provided from dense image matching or range sensors (aerial or terrestrial) is used to drive the edge extraction and the matching process. In this way the geometric object boundaries, useful in speeding up the drawing production or the building modelling are determined. This approach has been already presented in (Nex, 2010) where an exhaustive description of the algorithm is reported. In the following sub-sections the algorithm steps are briefly described (Figure 1), while the improvements developed in the edge extraction phase will be presented in the following section. The entire method, which starts from a dense and accurate point cloud and a set of images, is divided in blocks with concatenated processing steps.

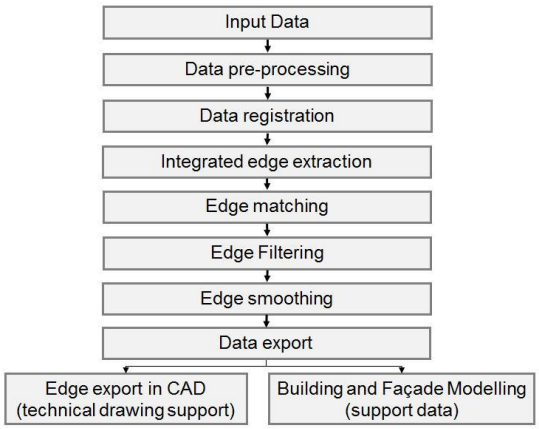


Figure 1. Algorithm workflow of the process.

- **Data acquisition.** Both in terrestrial and aerial applications the multi-image approach is mandatory as it allows to improve the quality and the reliability of the image matching results (Zhang, 2005; Nex, 2010). In terrestrial applications, a set of convergent images or acquired according to an *ad hoc* network geometry is suggested. In the aerial case, high overlaps between images and adjacent strips are mandatory.
High image resolution is always requested too: the extraction of building boundaries from images is complicated if low resolution images are used, as edges are usually blurred and irregular. The most central image of each set of images is considered as reference in the following matching process: when image sequences are available, different reference images are chosen in a proper way to assure a multi-view geometry.
- **Data pre-processing.** In order to improve the edge extraction, a non-linear Edge Preserving Smoothing (EPS) filter is applied to smooth the little radiometric variations but preserving and enhancing the main geometric discontents in the image. The boundaries of man-made objects are thus sharpened and smoothed deleting little radiometric changes that usually affect negatively the extraction of such elements. Moreover, in order to improve the radiometric contents of all the images and achieve better matching results, the Wallis filter is applied.
- **Data registration.** Image orientation can be performed using different approaches (Barazzetti et al., 2011, etc.). Images and range data have to be already oriented in the same reference system. This process is not necessary if the available point cloud is generated with image matching techniques. Otherwise, the point cloud need to be co-registered in the photogrammetric reference system by means of a spatial roto-translation.
- **Integrated edge extraction.** The integrated edge extraction is performed on a reference image, as described more in detail in following section. The final goal is to define only some "dominant" points able to provide a good approximation of the edge shape for its reconstruction in the matching process. The dominant points are recorded and linked by straight edges. The edges are extracted only in the regions of interest while areas where mismatches and blunders could occur are excluded with a manually masking approach.
- **Edge matching between images.** A multi-image matching approach, divided in three steps, is used. The first step is a modification of the Multi-Image Geometrically Constrained Cross Correlation (MIGC^3) proposed in (Zhang, 2005). Using a MIGC^3, the dominants points of each edge are matched in all the images in order to reconstruct the breakline positions in 3D (object space). The images are preliminarily undistorted (using the camera calibration data) in order to ease them into a central perspective and speed up the following processes. The MIGC^3 is able to match a high percentage of the extracted dominant point. Nevertheless, more than one reliable homologous point can be possible if only high cross correlation values are considered. A relational matching technique has been developed in order to solve these ambiguous matches and to improve the rate of the successfully matched points by means of a probability relaxation (Christmas et al., 1995). The method uses the already matched dominants points as "anchors" and defines, in an iterative way, the more suitable match between candidates imposing a smoothing constraint. Finally, a Multi-Photo Least Square Matching (Baltsavias, 1991) with the epipolar constraint is performed for each extracted point to improve the accuracy up to a sub-pixel dimension.
- **Edge filtering.** Once a set of 3D edges has been created, possible blunders are deleted using a filter that considers the reciprocal point positions on the same edge: the position of a point is "predicted" considering the neighbouring dominant points of the edge and then the difference between the predicted and real position of the point is evaluated. If the difference value is higher than a threshold, the point is delete. This filter works well if the blunders are isolated from each other. Then, more robust filter can be used to correct the edges when several blunders are close together: this algorithm uses the point cloud to verify the correctness of each dominant point: when it is farther than a threshold from the point cloud, the point is deleted.
- **Edge smoothing.** The edges extracted by the image matching algorithm are random noise affected and they cannot be directly used in the drawing production or in the segmentation process. For this reason, a smoothing is needed in order to define a regular shape of the object, easing the edges in lines and curves. The great majority of edges in both close range and aerial applications can be classified in sets of lines and second order curves. Therefore, each edge must be split in different basic entities that describe its linear or curved parts separately and, each separate basic entity is simplified in lines and curves fitting the dominant point information with a robust least square approach.
- **Edge exporting.** Geometric edges are exported in CAD environments in order to give a good preliminary data for the graphic drawing realization of the survey or to be used as additional information in the segmentation and modelling processes.

2.1 Integrated edge extraction

The completeness of the edge extraction process depends on several factors such as illumination condition, image resolution and typology of object. According to these factors, the edge extraction can produce fragmented or incomplete edges that reduce the quality of the achieved information.

Figure 2. Extracted edges using different Canny's thresholds.

For this reason, using a standard edge extractor on an image, incomplete results are usually produced. This problem can be partially solved decreasing the thresholds of the edge extractor (i.e. Canny operator), but this brings to an increase in the number of edges due to simple radiometric variations or shadows (Figure 2). The integrated edge extraction considers the range and image information in order to improve the continuity, length and reliability of the reconstructed edges in correspondence or in the closeness of only geometric discontinuities. As the conditions are very different in terrestrial and aerial acquisitions and the typology of objects to be detected is different too, two methods are used. In terrestrial applications, the goal is the detection of the geometric discontinuities of different shapes on the façade, while in aerial applications the building outlines and rooftops have to be detected. In both cases, the use of a model based approach is not flexible enough to correctly describe all the geometric discontinuities of a generic object. The common idea of these approaches is to reduce the regions where the edge extraction is performed using the range information. Edges are detected on the range data and then they are projected on the image defining a region of interest. In this way, the edges due to radiometric variations or shadows are almost completely ignored and only regions of the image in correspondence of geometric discontinuity are considered.

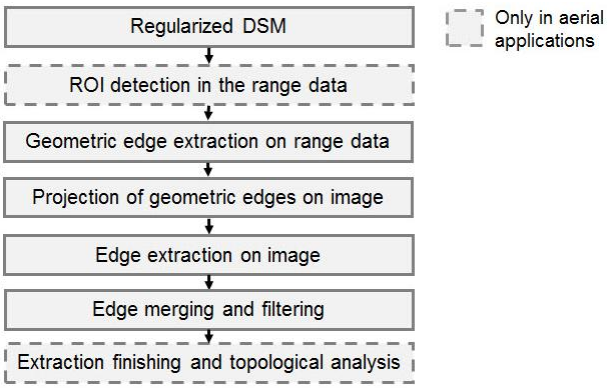


Figure 3. Integrated edge extraction workflow.

The algorithm works according to several steps (Figure 3), some of them performed only in aerial applications:

- **Regularized DSM**. The available range information is firstly regularized in order to define a depth image of the scene simplifying the successive computation steps. The regularization procedure, which considers the range information as 2.5D data, depends on the range data density and detail dimensions to be surveyed.

- **ROI detection in the range data**. Aerial acquisitions over a urban area usually considers 3 main classes of objects: ground (bare soil, grass, road, sidewalk, etc.), buildings and vegetation (i.e. trees). For this reason, the classification of building regions from range and image data is performed before the outlines and rooftop extraction. Simple RGB images and "single return" DSM (generated by LiDAR or image matching techniques) are used. To do that, different steps are performed. Off-ground data is formerly extracted from the data using the regularized range data (Figure 4). The simple assumption is that the height of a ground is lower than the heights of neighbouring non-ground points; the ground filtering is performed through an iterative regular grid filtering. This approach consider two different problems: (i) the ground height variations over a big region patch and (ii) the presence of big dimensions buildings that avoid to determine the ground height if a too little DSM patch is considered. For these reasons, the ground height is iteratively computed on different DSM patch dimensions.
The building terrain is then extracted from the off-ground data filtering the vegetation. To do that, the local height variability of the points is considered, as this value is higher in correspondence of vegetated areas than over manmade objects.

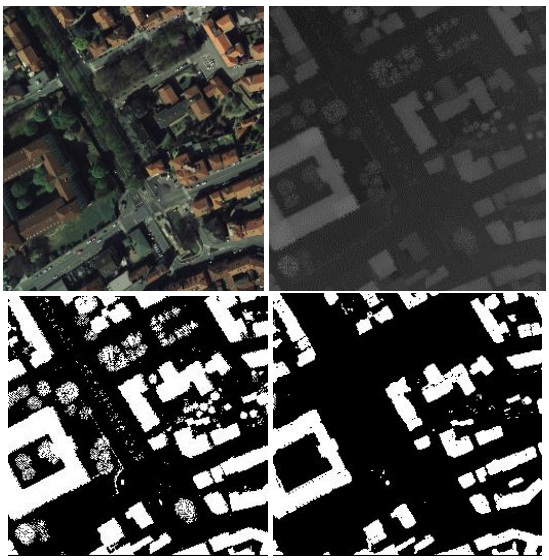


Figure 4. RGB image (upper left), regularized DSM (upper right), off-ground area (bottom left) and building (bottom right).

- **Geometric edge extraction on range data**. The geometric breaklines are extracted from the regularized range data (depth image) considering the gradients and the local curvature between adjacent pixels. This process is repeated both for each point of the depth image and the final result is a map of the geometric breaklines from the range data (Figure 5).

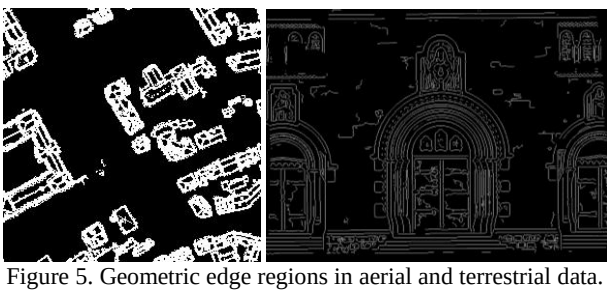
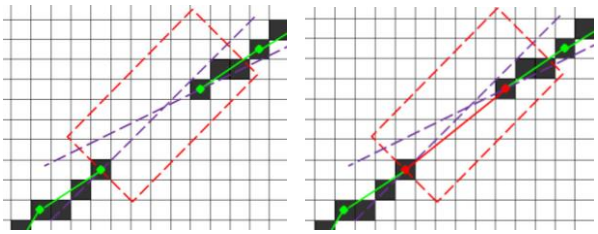


Figure 5. Geometric edge regions in aerial and terrestrial data.

- **Projection of geometric edges on image**. The breaklines detected on the depth image can be projected onto the oriented images. In this way, an interest region is defined in the image space and the edge extraction can be performed only in this area.

Figure 6. Geometric edge regions projected into the images.

- **Edge extraction on image**. The Canny's operator thresholds are stressed in order to preserve the edge continuity. In particular the lower threshold of the Canny's algorithm is further reduced in order to preserve the continuity between stronger edge points. The edges extracted with the Canny's operator are approximated considering each point where the edge changes its curvature as a dominant point, and linking with a straight line the points comprised between two adjacent dominant points.
- **Edge merging and filtering**. Although the Canny's thresholds are used in an adaptive mode, several edges can be divided or incomplete, thus requiring an aggregation strategy, in particular for long edges. The direction of the edge in correspondence of the extreme is computed considering the last dominant points of the edge. According to this direction a bounding box is defined and the presence of other edge extremes is looked for. When another extreme is found and the direction of the other edge extreme is similar, the edge is merged. The edges can be successfully merged only if the missing part of the edge is limited to few pixels.



(a) (b)

Figure 7. Edge merging process: the red box defines the bounding box for the merging process; violet lines defines the edge direction in correspondence of the extremes (red points).

- **Extraction finishing and topological analysis**. Several geometric features of a building cannot be described by using their radiometric content. To improve the completeness of the achieved results a combination of the geometric information and of the already extracted edges is considered.

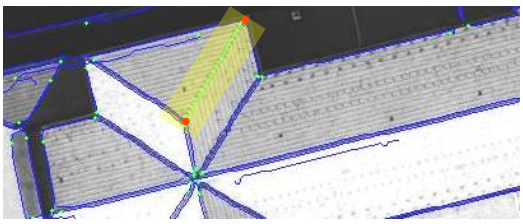


Figure 8. Reconstruction of missing parts of the roof.

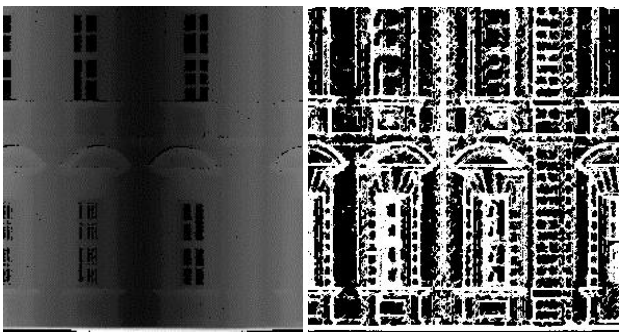
Roof outlines can be described by set of lines that are linked together in correspondence of knots (green points in Figure 8). Then, each outline on the range data must have a corresponding outline on the image. According to this, the implemented algorithm firstly defines if the roof description is complete comparing the range and the radiometric data: when a missing part is defined the knots in correspondence of this area (red points in Figure 8) are considered. Then, a topological analysis is performed in order to evaluate if these knots can be linked together. In this analysis some rules are considered: (i) the new edge and the geometric one (on the range data) must have approximately the same direction (yellow region in Figure 8); (ii) knots to be merged must belong to the same building; (iii) the new edge must be "consistent" with the other extracted edges and the other knots of the same roof (e.g. the new edge must not intersect other building edges, etc.). In this way, the correct between different points can be defined when several edges can be merged together (green line in Figure 8).

3. TESTS

In the following, some results on both terrestrial and aerial images will be presented. The terrestrial test was realized considering a set of convergent images of an historical building. The aerial test was performed on a dense urban area over the city of Torino (Italy).

3.1 Terrestrial test

This test was performed on 5 images (GSD ca 2 mm) acquired on a corner of an historical building of Torino (Italy). The LiDAR data was acquired with 2 cm spacing resolution and it was regularized with the same step, as shown in Figure 9 (a). In this way, the regions where the geometric discontinuities occurred were detected (Figure 9 (b)).



(a) (b)

Figure 9. Regularized DSM (a) and geometric breaklines (b).

Figure 10. Extracted edges on the reference image (about 96000 dominant points).

With such images the integrated edge extraction improved the completeness of the achieved information merging edges that can be divided or separated (Figure 10). The matching process allowed to determine the 3D position of the dominant points. These points were finally smoothed and eased in lines and curves (Figure 11).

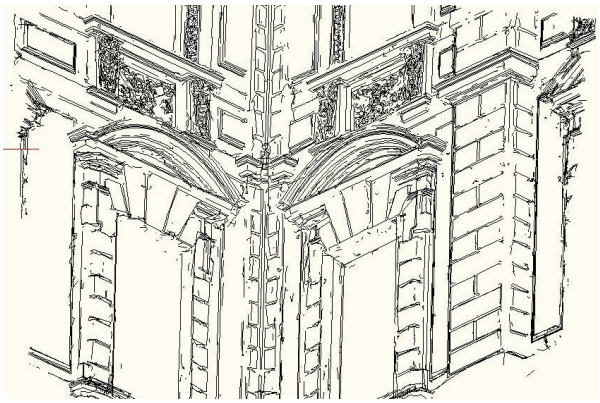
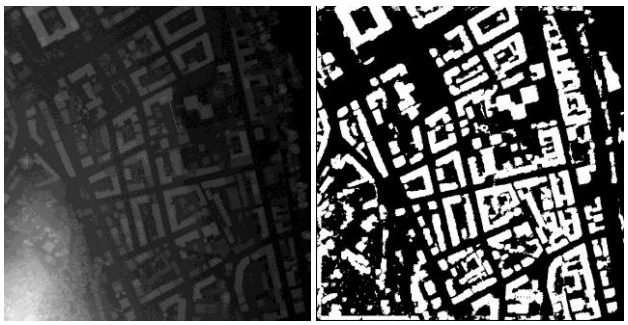


Figure 11. Achieved 3D results in the terrestrial test.

3.2 Aerial test

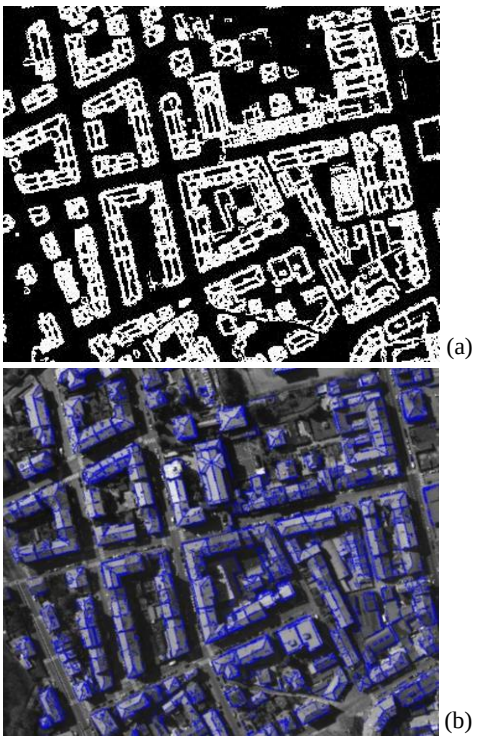
The dataset contains 6 aerial images (DMC Z/I Imaging-Hexagon Geosystems, GSD=12 cm) and 1m regular spacing DSM (single return) over a urban area of Torino (Italy). The test area (ca 0.5x0.5 km) is characterized by several high buildings, trees and variation of the ground height (Figure 12 (a)).
From the available DSM, the buildings were extracted and the ground and the trees were filtered according to the algorithm described in the former section. The result (Figure 12 (b)) is complete and correct even if some problems still remain where the vegetation is very close to buildings, as shown in the lower left part of the figure. In these conditions some trees are still classified as buildings.



(a) (b)

Figure 12. Regularizad DSM (a) and building regions (b).

The building rooftops and outlines were finally extracted on the same area (Figure 13). It can be noted that the results are complete and the areas of interest on the buildings are correctly detected. The regions detected in Figure 13 (a) were projected in the reference image to limit the edge extraction. The edges extracted on the image were finally filtered to delete the ones shorter than 20 pixels. The result after the whole process is shown in Figure 13 (b). Most of the building outlines and rooftops were correctly detected. Few lines are missing in correspondence of bad textured areas, but the number of useless points on the roofs and in proximity of buildings was reduced thanks to the filtering of the edges.



(a)

(b)

Figure 13. Roof outlines in the range data (a) and corresponding extracted edges on the images (b).

The merging of adjacent edges and the extraction finishing improved the quality of the result. The extracted edges were finally matched and smoothed. The final 3D result is shown in Figure 14.

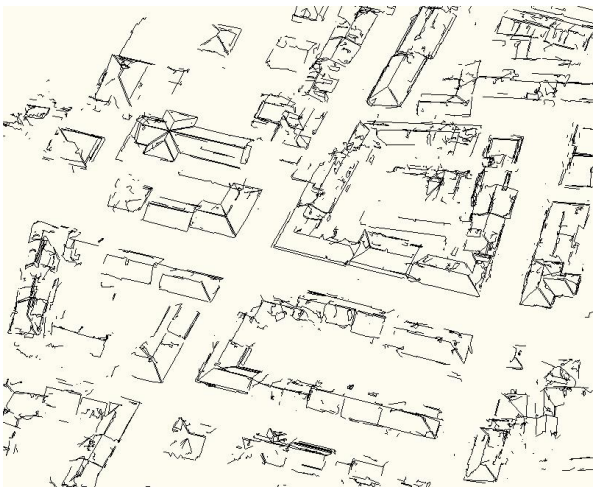


Figure 14. Achieved result in the aerial test.

The reported results, if compared to a former tests (Lingua et al. 2010) on the same area, show a significant improvement thanks to the integrated edge extraction that allowed to delete before the matching process the useless edges in correspondence of shadows and radiometric variations, increasing the completeness of the 3D product.

4. CONCLUSIONS AND FUTURE DEVELOPMENTS

In this paper an improved integrated approach for edge extraction from terrestrial and aerial images have been shown.

The method exploits both range and image information in order to provide for reliable and complete 3D edges useful for mapping and drawing purposes. The presented approach could complete the information provided by a LiDAR point cloud, that is generally unable to define the precise position of object breaklines and usually smooth the geometries of the surveyed objects. The proposed method gives also a complementary information with respect to dense matching algorithms as the extracted edges are concentrated in proximity of object breaklines and they are usually longer and more reliable.
The use of low resolution images or poor GSDs make the edge extraction more difficult and unreliable, delivering incomplete and very noisy edges. The integrated edge extraction presented in this paper partially solves this problem. The merging of adjacent edges is still insufficient to reconstruct and complete edge extraction process when the edge missing parts are long. For these reasons a good image resolution is recommended.
The aerial test has underlined the improvement that the integrated edge extraction provide to the final result. The feature extraction from the range data allows to detect the buildings outlines and rooftops in an almost complete way, reducing and easing the following edge extraction on the image. Then the edges are successfully filtered exploiting their geometric properties, i.e. their straightness and length. At the end of this process, approximately the 90% of building outlines has been detected on the images. Finally, the matching process and the smoothing allow to deliver more complete 3D edges.
The topological analysis has shown to be strongly influenced by the resolution of the DSM: 1m spacing DSMs are not sufficient to "drive" the reconstruction of the edge in a complete and satisfying way as their information is too rough to be used on high resolution images. On the other hand, the use of only radiometric edges positions is insufficient to reconstruct the roof geometry in a proper way.

REFERENCES

Alshawabkeh, Y., 2006. Integration of laser scanning and photogrammetry for cultural heritage. PhD Thesis Institut für Photogrammetrie der Universität Stuttgart.

Awrangjeb, M., Ravanbakhsh, M., Fraser, C.S., 2010. Automatic detection of residential buildings using LiDAR data and multispectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing 65*, pp. 457-467.

Baltsavias, E., 1991. Multiphoto geometrically constrained matching, PhD Thesis, ETH Zurich, Switzerland.

Barazzetti, L., Remondino, F., Scaioni, M., 2011. Automated and accurate orientation of complex image sequences. In: *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences XXXVIII (5/W16)* - 4th International Workshop 3D-ARCH 2011, CD-ROM.

Christmas, W. J., Kittler, J., Petrou, M., 1995. Structural Matching in Computer Vision Using Probabilistic Relaxation. *PAMI 17(8)*, pp. 749-764.

Demir, N., Poli, D., Baltsavias, E., 2009. Detection of buildings at airport sites using images & LiDAR data and a combination of various methods. In: *International Archives of Photogrammetry and Remote Sensing and Spatial Information Sciences*, XXXVIII(3/W4), pp. 71-77.

Gehrke, S., Morin, K., Downey, M., Boehrer, N., Fuchs, T., "Semi-global matching: an alternative to LiDAR for DSM generation?, 2010. In: *International Archives of Photogrammetry and Remote Sensing and Spatial Information Sciences*, Canadian Geomatics Conference XXXVIII(1).

Habib, A.F., Zhai, R., Kim, C., 2010. Generation of complex polyhedral building models by integrating stereo aerial imagery and LiDAR data. *Photogrammetric Engineering & Remote Sensing 76*, pp. 609-623.

Habib, A. F., Chang, Y-C., Lee D. C., 2009. Occlusion-based methodology for the classification of LiDAR data. *Photogrammetric Engineering & Remote Sensing 75(6)*, 703-712.

Hiep, V.,H., Keriven R., Labatut, P., Pons J-P., 2009. Towards high resolution multi-view stereo. Proceedings of: Computer Vision and Pattern Recognition, pp. 1430-1437.

Hirschmüller, H., 2008. Stereo processing by semi-global matching and mutual information, In: *IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2)*, pp. 328-341.

Lingua, A., Nex, F., Rinaudo, F., 2010. Integration of airborne laser scanner and multi-image techniques for map production. Proceedings of: SPIE Remote Sensing 7831, pp. 14-28.

Nex, F., 2010. Multi-image matching and LiDAR data new integration approach, PhD Thesis, Politecnico di Torino, Torino.

Paparoditis, N., Souchon, J-P. Martinoty, G., Pierrot-Deseilligny, M., 2006. High-end aerial digital cameras and their impact on the automation and quality of the production workflow. In: *ISPRS Journal of Photogrammetry & Remote Sensing 60*, pp. 400-412.

Pu, S., Vosselman, G., 2009. Knowledge based reconstruction of building models from terrestrial laser scanning data. *ISPRS Journal of Photogrammetry and Remote Sensing 64*, pp. 575-584.

Remondino, F., El-Hakim, S., Gruen, A., Zhang, L., 2008. Turning images into 3D models. In: *IEEE Signal Processing Magazine 25(4)*, pp. 55-64.

Sampath, A., J. Shan, 2008. Building Reconstruction from Airborne LiDAR Data Based on Clustering Analysis. In: *International Archives of Photogrammetry and Remote Sensing and Spatial Information Sciences*, XXXVII. Part B3a. Beijing 2008.

Zebedin, L., Klaus, A., Gruber-Geymayer, B., Karner, K., 2006. Towards 3D map generation from digital aerial images. *ISPRS Journal of Photogrammetry & Remote Sensing 60*, pp. 413-427.

Zhang, L., 2005. Automatic Digital Surface Model (DSM) generation from linear array images. PhD Thesis, ETH Zurich, No. 16078, IGP Mitteilung N. 90.

ACKNOWLEDGEMENTS

This work was funded by the 7° Framework Programme 2007-2013 under the name "CIEM Project" (co-founded by Marie-Curie Actions 7° P.Q. - PCOFOUND- GA-2008-226070, acronym "Trentino Project").

ESTIMATION OF SOLAR RADIATION ON BUILDING ROOFS IN MOUNTAINOUS AREAS

G. Agugiaro [a,*], F. Remondino [a], G. Stevanato [b], R. De Filippi [c], C. Furlanello [c]

[a] 3D Optical Metrology Unit, Fondazione Bruno Kessler, Trento, Italy
{agugiaro, remondino}@fbk.eu, http://3dom.fbk.eu
[b] Dept. of Architecture, Urban Modelling and Surveying, University of Padova, Italy
giulio.stevanato@unipd.it
[c] Predictive Models for Biomedicine & Environment Unit, Fondazione Bruno Kessler, Trento, Italy
{defilippi, furlan}@fbk.eu, http://mpba.fbk.eu

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: Photovoltaic potential, 3D buildings, Data integration, GRASS GIS, Photogrammetry, LiDAR, Terrain modelling

ABSTRACT:

The aim of this study is estimating solar radiation on building roofs in complex mountain landscape areas. A multi-scale solar radiation estimation methodology is proposed that combines 3D data ranging from regional scale to the architectural one. Both the terrain and the nearby building shadowing effects are considered. The approach is modular and several alternative roof models, obtained by surveying and modelling techniques at varying level of detail, can be embedded in a DTM, e.g. that of an Alpine valley surrounded by mountains. The solar radiation maps obtained from raster models at different resolutions are compared and evaluated in order to obtain information regarding the benefits and disadvantages tied to each roof modelling approach. The solar radiation estimation is performed within the open-source GRASS GIS environment using r.sun and its ancillary modules.

1. INTRODUCTION

The rising energy costs and the need to reduce carbon-dioxide emissions are intensifying research efforts for alternative, renewable (and sustainable) energy sources.
Solar technology is one of the natural choices for on-site generation as the energy coming from the sun is captured by solar panels and transformed into heating or, by means of photovoltaic (PV) systems, into electricity. The use of solar technologies is growing worldwide: large scale solar radiation maps (e.g. SOLEMI [1], SoDa [2], PVGIS [3]) are already published on-line; at urban scale some municipalities are starting to build city solar atlases (e.g. Hamburg [4], Berlin [6]) in order to increase or create the demand of photovoltaic and thermal panels (Ludwig and McKinley, 2010). The identification of suitable surfaces in urban areas plays therefore an important role both for the private investor and the public local community. Due to the complexity of this task, quality of solar radiation predictive models, as well as quality and quantity of their input data are pivotal to optimally exploit the advantages of solar panel systems. Indeed they need to be properly located and oriented in the environment to meet the required specifications (insolation time, area orientation, panel type, characteristics of power network, etc.).
Among all factors influencing a correct estimation of the incoming solar radiation, it is crucial to consider shadowing effects due to topography (presence of hills/mountains) or shadows cast by nearby buildings, vegetation or other objects found in urban areas (Ike and Kurokawa, 2005).
The accuracy and Level of Detail (LoD) of the geometric models used to represent buildings (or their roofs) is also important, since complex geometries – like in case of dormers or chimneys – must be considered in order to produce accurate solar radiation maps. The need of geometric accuracy, on one hand, and the variety of scales to be considered – from architectural to regional –, on the other hand, tend to be diverging forces, so that often a compromise has to be found. In the following, two examples are given to exemplify the two main research paradigms, which the method described in this paper is trying to unify.

a) PVGIS (Photovoltaic Geographical Information System) provides a map-based inventory of solar energy resources and assessment of the electricity generation from PV systems in Europe, Africa and South-West Asia (Šúri et al., 2006). For the European subcontinent it delivers, among other products, the daily sum of global radiation and the theoretical optimum inclination angle of PV modules to maximize energy yield production: data are presented as raster layers at a resolution of 1×1 km.
b) Jochem et al. (2009) focus instead on the automatic identification, segmentation and analysis of roof shapes/facets from a dense LiDAR dataset (17 points/m^2) in a study area of 0.3 km^2. Transparent shadow values for nearby vegetation are introduced and the full 3D information of the point cloud is used for both solar potential assessment and modelling of shadowing effects of nearby objects. The shadowing effect due to terrain is not directly considered, but included through the use the CSI (clear sky index) as a model of cloud-cover effects.

The work presented in this paper sits probably between these two research directions: the first goal is to test the feasibility of a solar radiation estimation methodology based on geometric data ranging from regional to architectural scale. Several roof models, obtained by different surveying/modelling techniques and with varying LoD, are embedded in a DTM of an Alpine

* Corresponding author. Full address: 3DOM, FBK – Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo – Trento, Italy.

valley surrounded by mountains, so that both the terrain and the nearby building shadowing effects are considered. Additionally, the solar radiation maps obtained from the different geometric models are compared and evaluated in order to gain information on benefits/disadvantages tied to each model.

The solar radiation estimation is performed within the GRASS GIS environment (Neteler and Mitasova, 2007) using r.sun and its ancillary modules. The open-source algorithms implemented in r.sun by Hofierka and Šúri (2002) are well-known and tested in a variety of studies, e.g. in Kryza et al. (2010); Nguyen et al. (2009) use r.sun to compute insolation including temporal and spatial variation of albedo and solar photovoltaic yield. All steps from data acquisition and pre-processing to post-simulation are presented, whereby candidate lands for incoming solar farms projects are identified.

Hofierka and Kanuk (2009) discuss a methodology for the assessment of photovoltaic potential in urban areas using open-source solar radiation tools and a 3D city model implemented in a GIS. The test area extends over ca. 3.7 km^2, the solar radiation is calculated using the PVGIS estimations coupled with the city model roof geometries. A comparison with r.sun and using building models has highlighted discrepancies between the averaged values of PVGIS and the spatial variability of an urban environment due to the shadowing effects of nearby object.

2. TEST AREA AND DATA SOURCES

The test site for this study is located in Mattarello, an urban borough of Trento, the largest city in the Trentino-Alto Adige/Südtirol region (Northern Italy). The test site area lies few kilometres south of the city centre, on the eastern flank of the river Adige valley, and it is surrounded, mainly east and west, by the Alps, whereas the Adige valley stretches mainly north to south. The test area is approximately 1.5×2.1 km wide, it contains circa 1300 residential, industrial and commercial buildings, with varying sizes and geometry complexity. Building location varies from the valley plane to hill top or on the flank of the Alps (Figure 1).

The dataset was derived from heterogeneous data sources, although at different levels of detail and covering different extents. Most of the used data were provided by the Autonomous Province of Trento (PAT). The datasets consist of:

- a vector *cadastral map*, at nominal scale of 1:1000, containing all building footprints and some external attributes, including the average building height. Since roofs are not modelled in the cadastral map, flat roofs were rasterised using the average height value of the building. The footprints served also in the next analyses as surface unit for each building: although rasterisation can introduce errors in the estimation of the total irradiated roof size (e.g. due to roof overhangs), this approach provides a standardised common reference for all geometry models employed in this work.
- a raster-based *DSM* (and the resulting DTM), at a nominal scale of 1:10000, of the whole province, calculated from a LiDAR flight in 2006/7. Height accuracy for the original LiDAR data is given as σ_z=15 cm for the DSM, and σ_z =30 cm for the DTM. Both DSM and DTM are delivered already post-processed, geo-referenced and rasterised. For the study area, raster tiles of 2 km side can be downloaded at 1×1 m resolutions. The DTM of a larger area surrounding Mattarello, of approximately 16×18 km and at 1 m resolution, was used in the successive computation of the horizon maps, i.e. considering adjacent mountains (as high as 1900 m) and the resulting shadowing effect.
- nadir *aerial images* acquired from a helicopter in 2009 with a calibrated Nikon D3X camera equipped with a 50 mm lens. The images have an average ground sample distance (GSD) of circa 10 cm and were triangulated and geo-referenced in ERDAS/LPS using 6 GCPs measured with sub-decimetre accuracy by means of GNSS a receiver. From the aerial images, four different models were extracted: two DSMs, at 1 m and 25 cm resolution respectively, generated automatically with SAT-PP [6]; the exported vector 3Dfaces of 30 manually measured roofs, modelled in PhotoModeler [7], were rasterised to produce the last two datasets, at 1 m and 25 cm resolution, respectively.

Six raster models of the study area were generated and used for the solar radiation estimation in the test area:

a) a raster of flat roofs derived from the cadastral map,
b) a raster with roofs extracted from the LiDAR-based normalised DSM overlapping the cadastral map;
c) a raster with roofs obtained from the automatic matching of aerial images;
d) a raster with roofs obtained using PhotoModeler;
e) an analogous raster to c) but rasterised at higher grid resolution;
f) an analogous raster to d) but rasterised at higher grid resolution.

Rasters a), b), c) and d) are at 1 m resolution; e) and f) at 25 cm resolution. All rasters were embedded in the DTM (embedding of the rasters at 25 cm resolution required DTM oversampling). All raster models were compared with each other in order to identify and exclude those buildings which have been changed, demolished or built in the time interval between different data acquisitions. A detail view of all six rasters is given in Figure 2.

Figure 1 – Aerial view (Google Earth) of the Adige valley in the Alps, with the location of the test site: Mattarello, Trento, Italy.

3. ESTIMATION OF THE SOLAR RADIATION

Šúri and Hofierka (2004) identified three main factors determining the interaction of the solar radiation with the Earth's atmosphere and surface. The first one depends on the terrestrial geometry, i.e. the rotation and revolution of our

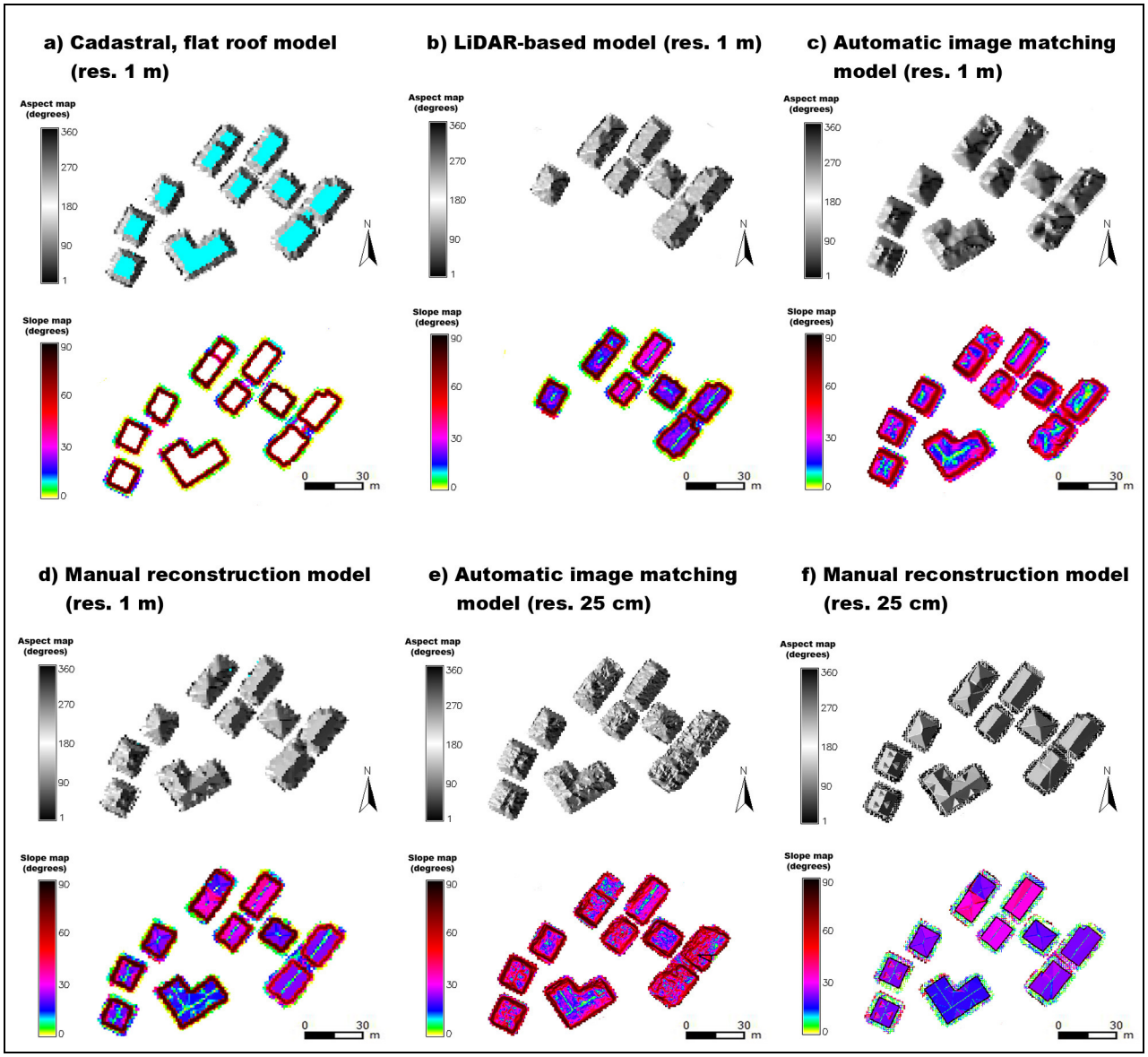


Figure 2 – Raster maps obtained from different data for building roofs. Aspect and slope maps of a group of buildings in the test area, obtained from: a) cadastral maps (flat roofs), b) LiDAR-based DSM, c) automatic matching of aerial images, d) manual reconstruction from aerial images, e) automatic matching of aerial images (rasterised at grid resolution 25 cm), f) manual reconstruction from aerial images (rasterised at grid resolution 25 cm). Aspect maps are classified starting from east, counter clockwise (north=90°, west=180°, south=270°), areas in cyan are horizontal. Some roofs are missing in b) since not yet built at the time of the LiDAR flight. Terrain data (except a thin buffer zone) has been masked out in order to facilitate visualisation.

planet around the sun, which determines the available extra-terrestrial radiation based on solar position above horizon. Secondly, the terrain surface, i.e. the slope, the aspect and shadowing effects of the surrounding terrain features can modify the radiation distribution to the Earth's surface. Finally, the atmosphere is composed by gases, clouds, solid and liquid particles, which all lead to a certain attenuation in terms of global radiation.

In GRASS GIS, the r.sun module allows to model all above mentioned factors, although at different levels of accuracy.

The geometric factors (astronomic and terrestrial ones) can be modelled quite efficiently, while the atmospheric attenuation can be handled only with a certain level of accuracy. More specifically, r.sun computes direct, diffuse and ground reflected solar radiation maps for a given day, latitude, surface and atmospheric conditions, using built-in solar parameters (e.g. time of sunrise and sunset, declination, extra-terrestrial irradiance, daylight length). The model computes radiation for the clear sky conditions, thus it does not take into consideration the spatial and temporal variation of clouds. Average monthly values of the air turbidity coefficients (Linke data) can be provided as a single value or as input raster maps: in this study the Linke data were obtained from the SoDa site [2] as a global dataset, then reprojected and resampled to the local coordinate system.

The shadowing effect of the topography can also be incorporated. It is achieved in two ways: it is calculated directly by r.sun from the digital elevation model or, alternatively, rasters of the horizon height are used. Horizon maps are pre-computed only once, so their use speeds up r.sun operations considerably (this second approach is preferable in case of multiple simulations). The GRASS module r.horizon is

used to iteratively compute horizon maps for a given area, in that *n* maps are created for *n* directions: for each cell, the horizon height angle is stored for the given direction in a map. Moreover, although horizon raster are needed only for the study area (i.e. approximately 1×2 km), r.horizon allows to extend the computation area to larger parts of the surrounding DTM, thus including shadow-casting mountains around the test site. In this study, 24 horizon maps were computed for each geometric model, thus at 15° intervals. This was the most demanding step in the whole pipeline: all computations were carried out on a 3 GHz dual-core machine, with 8 GB of RAM and running a 64 bit version of Linux and GRASS GIS 64 bit. Computation of 24 horizon maps at 1 m grid resolution took approximately 10 hours, while the time needed for a 25 cm model was about 3.5 days.
Once the elevation model, its aspect and slope maps, the Linke turbidity and the horizon maps are prepared, solar radiation can be computed using r.sun. More specifically, several radiation maps were obtained in this work: for each geometric model, direct, diffuse and global radiation maps were calculated, yielding the average radiation value (in Wh/m^2/d) for each month. The yearly average value was also calculated.

3.1 Calibration of r.sun

In order to calibrate the global radiation model, real data collected in the past 10 years (2001-2010) from a pyranometer, located on the roof of an industrial building next to the study area were used. The availability of solar radiation flux density values (W/m^2) sampled every 15 minutes allowed to compute daily clear sky global radiation (GR) values at the pyranometer position using r.sun with the following inputs: a) the DSM of the area, including the shadow-casting nearby mountains, b) horizons maps with an angle step of 20°, c) Linke data previously described. Furthermore, the pyranometer data were aggregated in order to obtain a daily global radiation values.
Since cloud cover information is not available, the daily maximum values (MV) over the ten years observation interval were used as observed values for clear sky solar irradiation.
By means of the normalised mean bias (NMB) index, under- or overestimations of the model can be quantified as follows on a monthly basis (N=days per month):

$$NMB = \frac{\sum_{i=1}^{N}(MV_i - GR_i)}{\sum_{i=1}^{N} GR_i} \qquad (1)$$

The NMB index was computed along the 10 years for December, March and July. The model was found to underestimate in December (-20.5%) and overestimate in July (12%), respectively. A good correspondence between estimated and observed values was found in March (0.2%).

3.2 Solar radiation and roof models

Multiple r.sun simulations were run stating from the six different rasters (see section 2) and the accompanying data (slope, aspect, horizon maps). Direct radiation, instead of global radiation, was chosen for these analyses, since it is mostly affected by the shadowing. On all resulting maps, yielding the monthly average direct solar radiation expressed in Wh/m^2/d, cell values were aggregated using the cadastral footprints as aggregation area. For each building, comparable minimum, maximum and average values of direct radiation were obtained, in order to compare the different simulation results.

An initial test was performed to check the influence of the shadowing effect: radiation maps were calculated from the LiDAR-based raster, without inclusion of the terrain shadowing effect. In general, the shadowing effect due to the topography led to lower values of solar radiation, as expected: on a year basis, differences between homologous roofs (with and without terrain shadowing) are up to 8% on some buildings; on monthly basis, differences can reach peaks of 38% during the winter months (December and January).
Taking the whole dataset into account for rasters of type a) and b) (see Figure 3), radiation results for the cadastral flat roofs are consistently higher than for LiDAR-based roofs by 4.0% on a year basis. The difference has a minimum in December (2.0%) and a maximum in June (4.5%). Results are plotted in the graph of Figure 3.

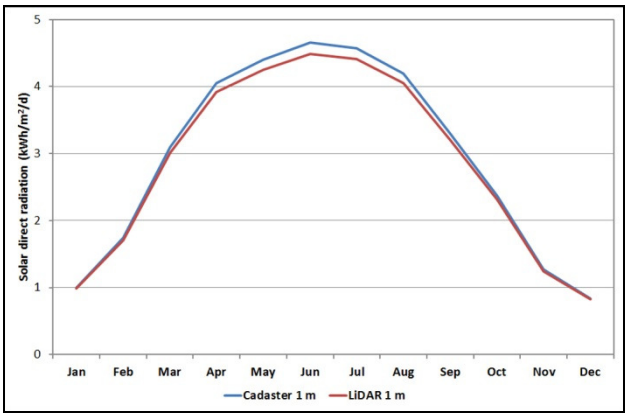


Figure 3 – Monthly solar radiation values from cadastral and LiDAR-based roof models, obtained averaging results from all 1300 buildings in the test area.

The analysis was replicated on the same dataset, distinguishing however between large industrial/commercial buildings (which tend to have a quite regular geometry and flat roofs) and smaller residential buildings, whose roof geometries may be more variable. A smaller difference of 1-2% on a year basis was found for non-residential buildings than for residential ones (4-6% on a year basis). In the latter case LiDAR-based roofs had peak differences of 13% per year. An example showing results from an industrial and a residential building is shown in Figure 4.

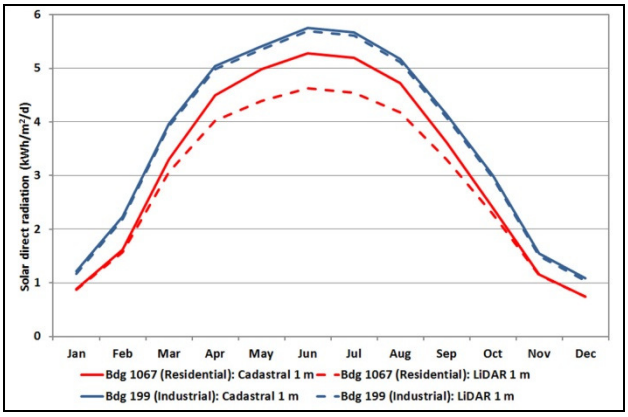


Figure 4 – Comparison between solar radiation results obtained for two different building classes. In case of an industrial building, cadastral and LiDAR-based roofs deliver similar results (average difference 1.9% on a year basis). For a residential building, results differ more (average difference

9.6% on a year basis, maximum difference in June, 12.4%, minimum difference in December, 0.1%).

For a subset of 21 residential buildings, rasters of all other types c), d), e) and f) were computed, thus including the effect of different methods and scales in the automatic matching and manual reconstruction process. A comparative analysis was performed by using at first only results from rasters at 1 m resolution, then including also those obtained at 25 cm resolution.
At 1 m spatial resolution, comparing automatic image matching to manual reconstruction, it was found that roof models from LiDAR data deliver fairly similar results, while a higher difference was found for cadastral flat roofs, as in the previous experiment. Taking the LiDAR-based roof results as reference, yearly average differences account for: -2.1% manual reconstruction models, 0.9% automatic matching models, and 6.9% cadastral roof models. The monthly average solar radiation values for all models are represented and plotted in the graph of Figure 5.

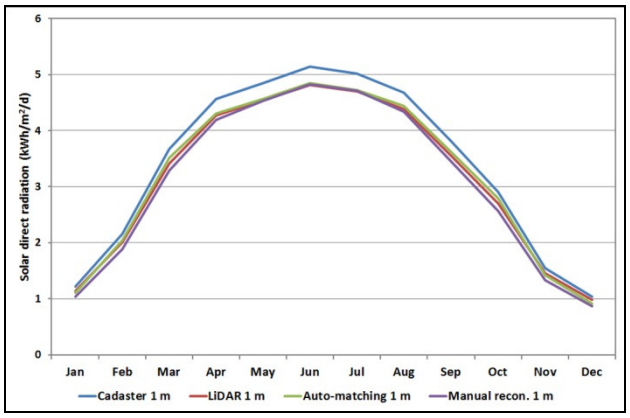


Figure 5 – Comparison between solar radiation results obtained from all rasters at 1 m resolution.

At 25 cm spatial resolution, results show that roofs modelled by automatic image matching yield the lowest solar estimate of radiation. Manually modelled roofs lead instead to values of solar radiation that, during the summer months, are closer to those of the cadastral flat roofs. Taking again the LiDAR-based roof results as reference, yearly average differences account for 1.8% in manual reconstruction models and -11.3% in automatic image matching models. In Figure 6 results from the 25 cm resolution rasters are plotted over the same data of Figure 5.

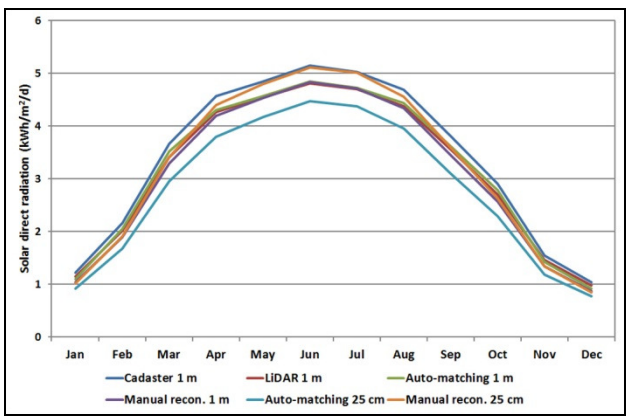


Figure 6 – Comparison between solar radiation results obtained from all rasters, at 1 m and 25 cm resolution. Data at 1 m resolution are the same as in Figure 5.

Although a subset of 21 roof models is still numerically too small if compared with the whole dataset of 1300 buildings – and therefore the following considerations must be taken with care and are subject to further testing –, comparative analyses suggest the following comments to the results obtained so far:

a) As long as solar radiation estimation are carried out on small to mid-size residential buildings, there are no substantial differences in the output among rasters at 1 m resolution models obtained from the LiDAR-based DSM, the automatic image matching and the manual modelling process. The cadastral flat roof models tend instead to deliver higher values of direct solar radiation. Therefore, the more time-consuming manual modelling could be avoided whenever other data-sources are already available at such resolution.
b) When using rasters at 25 cm resolution obtained from manually modelled roofs, results are again comparable with those at 1 m resolution (except flat roofs). The reason for higher values of solar radiation in the summer months is still subject of investigation, although one possible explanation could be the absence of chimneys in the PhotoModeler reconstructions: such a roof surface is therefore not affected by their shadowing effect.
c) When using rasters at 25 cm obtained from automatic image matching models, values of solar radiation are considerably lower than those obtained at 1 m. Although the reason is still under examination, this could be due to noise introduced by the auto-correlation algorithms on otherwise planar roof surfaces during the DSM extraction. This is best seen in Figure 7: the original roof facets are quite regular (there is only one chimney), nevertheless the resulting DSM is not correctly modelled and might be affected by "auto-shadowing".

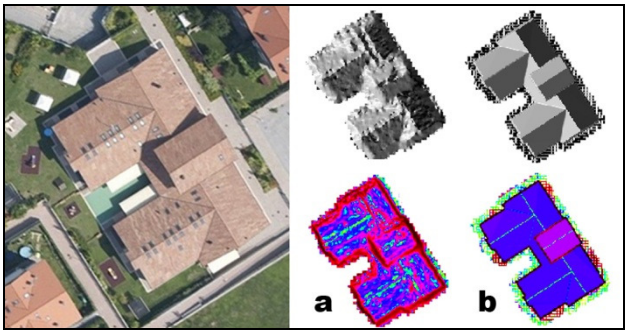


Figure 7 – Comparison between rasters at 25 cm resolution of the same building, presented in the left image: aspect (top) and slope (bottom) maps of the same building obtained from photogrammetry: automatic matching (a) and manual reconstruction (b). Legends for aspect and slope are as in Figure 2.

4. CONCLUSIONS AND OUTLOOK

In this work, different roof models have been rasterised and integrated onto a regularised DTM for the estimation of the solar radiation: flat roofs from cadastral datasets, LiDAR-based DSM roof surfaces, as well as models obtained from automatic and manual image matching. All models have their own peculiarities in terms of geometric resolution and accuracy and

deliver slightly different results. However, they all represent a step forward compared to the otherwise available solar radiation estimation models in terms of geometric accuracy for the Trento area: from the PVGIS raster based analyses at 1 km grid cell size, it now possible to perform solar radiance simulations on roofs at resolutions ranging from 1 m to 25 cm.
The test site has been chosen due to the availability of heterogeneous datasets, its complex topographic characteristics (i.e. a valley between high mountains) and the presence of already installed solar panels which guarantee real data to be used as reference for global solar radiation.
In the computation process, shadows cast by the surrounding mountains have been considered.
For every raster obtained from a distinct data source, monthly average solar direct radiation maps have been obtained, in order to perform a comparison among the results of homologous buildings. Radiation maps obtained with rasters at 1 m resolution show, in general, that results from the LiDAR-based roofs and those from automatic image matching and manual modelling deliver comparable results. The flat roof models from cadastral maps tend to provide higher estimations of solar radiation, however they can be used as well in case of big industrial or commercial buildings with fairly regular and planar roofs.
When it comes to the analyses of the 25-cm-resolution rasters, it must be primarily noted that more high-resolution models are needed to perform further tests. Nevertheless, the initial results indicate that the roof models obtained from automatic image matching do not lead to results similar to those from other models. More specifically, direct solar radiation values are generally underestimated by circa 10%.

Several are the future planned enhancements and possible extensions to the presented approach. First of all, no shadowing effect from nearby vegetation is yet integrated in the pipeline; however this could be achieved using data from the DSM, at least as long as no other better data sources are given.
In general, solar radiation results obtained so far lack a proper comparison with reference data of incoming solar energy. This is due to the fact that solar panel installations on private houses are generally mounted without any pyranometer. Often, even a data-logger to record the resulting electricity production is missing or sold optionally. Nevertheless, validation of the r.sun model could be carried out using data being logged from existing PV industrial installations next to the study area.
Moreover, r.sun already implements the possibility to further refine/reduce the clear-sky radiation by means of proper atmospheric correction coefficients (e.g. cloudiness), however these values derive from long-term meteorological measurements and must be provided separately for direct and diffuse radiation (Šúri and Hofierka, 2004).
All analyses so far are expressed in in $Wh/m^2/d$ and averaged over the whole footprint of the building. This is of course a strong simplification, since no segmentation has been carried out yet on the existing datasets to identify the distinct roof facets. On the other hand, this allows to deliver comparable results from heterogeneous datasets in a relatively fast and nearly completely automated way.
It must be noted that very dense and accurate geometric data is surely needed to model complex roof geometries (e.g. chimneys and dormers), otherwise the effort of high resolution analyses may not be worth.
It could be therefore interesting to test this methodology on an existing, detailed and already segmented city model (e.g. CityGML LoD 2 or 3) (Kolbe, 2009). Starting from such a spatio-semantic rich model could help the automation process to model the solar radiation (and thus PV potential) for all urban objects, providing furthermore local authorities with a powerful planning and information tool.

REFERENCES

Hofierka, J., Kanuk, J., 2009: *Assessment of photovoltaic potential in urban areas using open-source solar radiation tools.* In: Renewable Energy, Vol. 34(10), pp. 2206-2214.

Hofierka, J., Šúri, M., 2002: *The solar radiation model for Open source GIS: implementation and applications*. In: Manuscript submitted to the International GRASS users conference in Trento, Italy.

Ike, S., Kurokawa, K., 2005: *Photogrammetric estimation of shading impacts on photovoltaic systems.* In: 31st Photovoltaic Specialists Conference, IEEE, 3-7 Jan. 2005, pp. 1796-1799.

Jochem, A., Höfle, B., Hollaus, M., Rutzinger, M., 2009: *Object detection in airborne LIDAR data for improved solar radiation modeling in urban areas.* In: International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences. Paris, Vol. 38 (part 3/W8), pp. 1-6.

Kolbe, T, 2009: *Representing and exchanging 3D City Models with CityGML.* In: Proceedings of the 3rd International Workshop on 3D Geo-Information, Lecture Notes in Geoinformation & Cartography, Seoul, Korea, Springer Verlag, p. 20.

Kryza, M., Szymanowski, M., Migala, K., et al., 2010: *Spatial information on total solar radiation: Application and evaluation of the r.sun model for the Wedel Jarlsberg Land, Svalbard*. In: Polish Polar Research, Vol. 31(1), pp. 17-32.

Ludwig, D., McKinley L., 2010: *Solar Atlas of Berlin - Airborne Lidar in Renewable Energy Applications*. In: GIM International, Vol. 24 3/2010, pp.17-22.

Neteler, M., Mitasova, H., 2007: *Open Source GIS: A GRASS GIS Approach.* Third edition. Springer, New York, ISBN-10: 038735767X; ISBN-13: 978-0387357676.

Nguyen, H.T., Pearce, J.M., 2010: *Estimating potential photovoltaic yield with r.sun and the open source Geographical Resources Analysis Support System.* In: Solar Energy, Vol 84(5), pp. 831-843.

Šúri, M., Hofierka, J., 2004: *A new GIS-based solar radiation model and its application to photovoltaic assessments.* In: Transactions in GIS 8, pp. 175-190.

Šúri, M., Huld, T., Dunlop, E.D., Albuisson, M., Wald, L., 2006: *Online data and tools for estimation of solar electricity in Africa: the PVGIS approach*. In: Proceedings of the 21st European Photovoltaic Solar Energy Conference and Exhibition, 4-8 October 2006, Dresden, Germany.

REFERENCES FROM WEBSITES

1, http://www.solemi.com (Last visit: 29 June 2011)
2, http://www.soda-is.com (Last visit: 29 June 2011)
3, http://re.jrc.ec.europa.eu/pvgis (Last visit: 29 June 2011)
4, http://www.hamburgenergiesolar.de (Last visit: 29 June 2011)
5, http://www.businesslocationcenter.de/en/3d/A/i/1/seite0.jsp (Last visit: 29 June 2011)
6, http://www.4dixplorer.com/software_satpp.html (Last visit: 29 June 2011)
7, http://www.photomodeler.com (Last visit: 29 June 2011)

ACKNOWLEDGEMENTS

The presented work was partly supported by the ENERBUILD project within the Alpine Space Program (11-2-1-AT). The authors would also like to thank the Autonomous Province of Trento (PAT), which kindly provided most of the employed spatial data. Thanks are also due to Unifarm for providing solar radiation data logged by the pyranometer, as well as to Shamar Droghetti and Andrea Gobbi (FBK Trento) for valuable help during the data processing and model validation.

SMART FILTERING OF INTERFEROMETRIC PHASES FOR ENHANCING BUILDING RECONSTRUCTION

A. Thiele [a,b] *, C. Dubois [a], E. Cadario [b], S. Hinz [a]

[a] Karlsruhe Institute of Technology (KIT), Institute of Photogrammetry and Remote Sensing (IPF), 76131 Karlsruhe, Germany – (antje.thiele, stefan.hinz)@kit.edu, clemence.dubois@student.kit.edu
[b] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB), 76275 Ettlingen, Germany – (antje.thiele, erich.cadario)@iosb.fraunhofer.de

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: Remote Sensing, Building Reconstruction, High Resolution InSAR

ABSTRACT:

The current generation of space borne high resolution SAR sensors provides high spatial resolution as well as interferometric data within short time frames. This makes such data attractive for 3D information extraction. Especially, the operational configuration of TerraSAR-X and TanDEM-X opens up new perspectives for this kind of applications. Despite of this, the interferometric phases still suffer from considerable noise, so that filtering is mandatory to enhance building reconstruction.
In our previous work, we used conventional Multilook-filtering to smooth the phase signature. For large buildings acceptable filter results are shown, but signatures of small buildings and significant layover areas are destroyed by the use of large square windows. Such filters are especially inappropriate if building orientations are not aligned with the sensor flight direction. Hence, in this paper, we present modified InSAR phase filters to support 3D building reconstruction. The implementation focuses on two different strategies: on the one hand taking GIS information into account, in order to parameterize the filters accordingly, and on the other hand purely relying on the image data. The filters are tested on simulated interferometric phases and on real single-pass airborne InSAR data. Finally, filter properties are compared with current standard InSAR filters.

1. INTRODUCTION

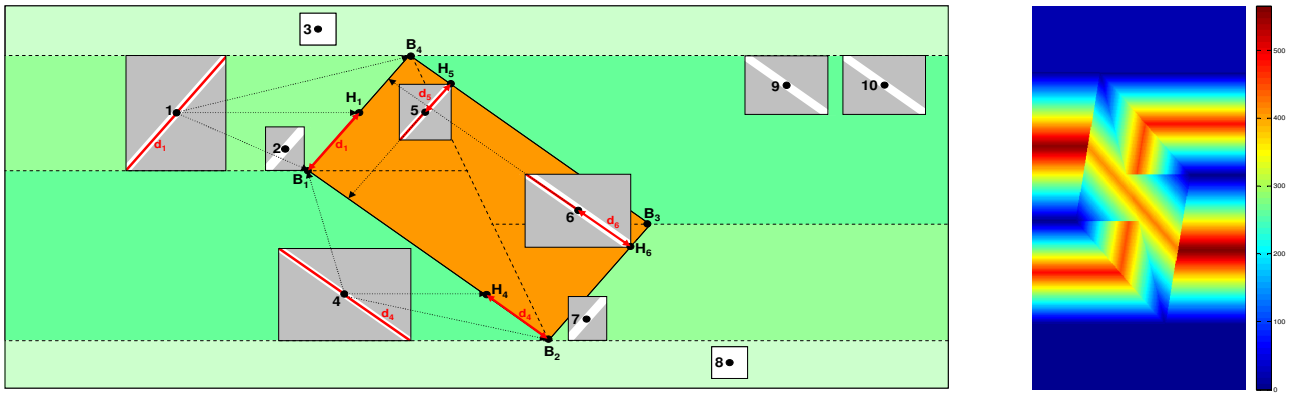
1.1 Motivation

In the last years the new generation of spaceborne high resolution SAR sensors such as TerraSAR-X, SAR-Lupe, Cosmo-SkyMed or RADARSAT-2 provides SAR images of meter resolution or even better in special spotlight modes, which open up the floor for many new applications. In particular, the development of methods to automatically derive detailed cartographic information of both rural and urban areas from this kind of data is a major issue driven by these missions.
Now, the newest SAR satellite sensor systems provide short repeat-pass or even highly coherent, single-pass interferometric data, which makes such data attractive for 3D information extraction. In particular the operational configuration of TerraSAR-X and TanDEM-X opens up new perspectives for this kind of applications.
Building detection from InSAR data presented in the literature was mainly based on a combined analysis of magnitude and interferometric height data (Bolter 2001, Soergel et al. 2003, and Thiele et al. 2007a). The utilization of the magnitude signature focused mainly on analyzing layover and shadow areas; and the analysis of the interferometric heights was mostly restricted to mean height calculation within an estimated building footprint. Yet, in high resolution InSAR data, even the shape of the interferometric phase profile at building locations contains valuable information (Thiele et al. 2007b).
A concept of exploiting this information for 3D building reconstruction was already presented in Thiele et al. 2010. Our reconstruction utilizes available 2D GIS information (building footprint) to simulate interferometric phase signatures of buildings. An iterative process of assessing real and simulated InSAR phases, of updating building model and of repeating simulation is set up to achieve best reconstruction results.
First tests revealed, that the assessment step between the simulated and real measured InSAR phases is the most crucial point. The reliability of the results depends on sensor configuration (e.g., baseline length), on quality of InSAR data (e.g., single-pass, repeat-pass), on used post-processing (e.g., filtering) and on matching between building model and real building. Hence, we now focus on the optimization of the smoothing of noisy InSAR phases using smart filtering.
Conventional Multilook-filtering (Lee et al. 1994) yields acceptable results when applied to large homogenous areas, but characteristic phase signatures, which appear also within layover areas of buildings, or signatures of smaller buildings were destroyed in particular by the use of large filtering windows. Similarly, such approaches are inappropriate if building orientations are not aligned with the sensor flight direction.

1.2 Related Work

InSAR filters like proposed in Goldstein et al. 1998 and Baran et al. 2003 investigate the frequency spectrum of an InSAR patch to reduce high frequency noise in the InSAR phases. Additionally, a weighting is applied depending on scene properties or local coherence values. Another filter approach presented in Tupin 2011 analyses coherence between so-called non-local image areas to de-noise image patches by a regularization-based method. This study is motivated by speckle reduction in SAR intensity data, but promising results are also shown on InSAR phase data.

* Corresponding author.



a b

Figure 1. Adaptive phase filtering: a) schema of filter approach and b) 2D histogram of adaptive filter size [pixel number]

Improvements on conventional Multilook-filtering are described in Reeves et al. 1999 to increase smoothing while avoiding the elimination of significant local features. For this reason interferometric amplitudes are used to specify filter weights, which are finally determined by utilizing a Monte-Carlo scheme. Approaches on adaptive complex Multilooking motivated by coherence filtering are published by Ciuc et al. 2002 and Vasile et al. 2004. Both start with a statistical analysis of SAR amplitude data, followed by a two-stage region growing on the same data. Subsequently, a complex averaging in the segmented adaptive neighbourhood is achieved by considering different weighting functions.
A windowing and segmentation independent approach on phase noise modelling and reduction is described in López-Martínez et al. 2002. The filtering of phase noise is accomplished by a local analysis in the wavelet domain, which enables high computational efficiency. Phase filtering by investigating morphological operators was presented in Rejichi et al. 2010. Based on gradient estimation in interferometric phases an alternate sequence of opening and closing operators is applied. Lee et al. 1998 implemented an adaptive orientation filter based on local noise level in the phase data. Sixteen orientations are considered and the local weighting is characterized by locale coherence, number of looks, and locale variance. A related approach is given in Bo et al. 1999 by extending the number of orientation masks. Two different values of line thickness are defined to preserve small signature details. The filtering and weighting is characterized by two options median or mean, similar to Lee et al. 1998.

Our implementations focus on two different strategies: on the one hand taking GIS information into account, and on the other hand, relying purely on image data without using additional GIS information. The paper is organized as follows: In Sect. 2, we outline the modifications and development for our smart filtering schemes. Test data are introduced in Sect. 3, followed by the visual and numerical validation of the filter results in Sect. 4. Conclusion and outlook are eventually given in Sect. 5.

2. INTERFEROMETRIC PHASE FILTERING

The investigated filters as well our adapted and new filters can be subdivided into three different groups characterized by the chosen type of filter window and by the introduction of prior information.
The first group contains the patch filter applied in the frequency domain on complex data (Goldstein-filter and Baran-filter) or in the spatial domain on phase data (coherence weighted Mean-filter and Median-filter).
The second group utilizes filter masks of different orientation to enable an adaptive averaging depending on local noise level or additional information. An adapted Lee-filter (org. Lee-filter without local unwrapping) and GIS-information (only orientation) related fixed-window-filter (GIS FWF) are considered.
The third group – adaptive-window filters – integrate building footprints. For these GIS-information related adaptive-window-filters (GIS AWF), coherence weighted averaging (mean+coherence) and median searching (median) are implemented. The parameterization of the adaptive filter window is visualised in Figure 1a. Considering a single building, the surrounding and the building (corner points B_1, B_2, B_3, B_4) are divided in three zones (dashed lines and different green colours), where different adaptive filter windows are applied. For the first area containing the example point 3 and 8 only a 5 x 5 square window is used. Exemplary filter windows for the second zone are given at the points 1, 2, 5, 7 and for the third at points 4, 6, 9, 10. The respective orientation in the filter window (considered pixels white marked) depends on the orientation of corresponding (closest) building side. The length of filter diagonals of pixels outside building footprint is defined by the doubled distance between H_i and the nearest building corner (see d_1 and d_4). For pixels inside the building, the perpendicular distance to the opposite building side is used (see d_5 and d_6). In addition, a map indicating the number of pixels taken into account for averaging and median search is given in Figure 1b.
The above filtering schemes are applied to the wrapped phases. Future work will include the implementation of local phase unwrapping during filtering. One solution is the calculation of a mean phase level in a neighbourhood like Lee et al. 1998. For this approach, an additional integration of prior information (e.g., simulated InSAR phase) might be even more useful.

3. TEST DATA

The different groups of filters are applied on simulated data and real InSAR data sets showing the same industrial building with a size of approx. 60 m width and approx. 100 m length. The investigation on simulated phase data allows to investigating the different filtering behaviour in detail, especially for significant signature changes, e.g., between layover area and building roof area. Due to the fact that the utilized phase simulation approach delivers no complex data but only phases, some current phase filters can only be tested on real InSAR data. To this end, single-pass airborne SAR imagery was used.

3.1 Simulated Interferometric Phase Data

For the simulation of interferometric phase, we use an adapted version of the approach presented in (Thiele et al. 2007b). A 3D building model and some SAR sensor parameters are needed as input. The building footprint is expected to be available from GIS data (Figure 2a,b) and the sensor parameters are chosen similar to the investigated real InSAR data. In contrast to earlier work, we compute not only a single phase profile per building hypothesis is calculated but the full interferometric phase signature of the building. In the simulation step we consider the fact that, especially at building locations, a mixture of several scattering effects can contribute to the measured interferometric phase within a single resolution cell. Hence, phases in the layover area show the so-called front porch shape mentioned in the literature. An example of the simulation result is given in Figure 2c.
For testing the different filter approaches, the simulated phases have to be corrupted by noise. The well-known noise distributions of the following instances are used contributing to the noisy (simulated) phase $\Delta\varphi_{noisy}$:

$$\Delta\varphi_{noisy} = \Delta\varphi_{sim} + \Delta\varphi_{thermal} + \Delta\varphi_{coherent} + \Delta\varphi_{shadow} \quad (1)$$

where $\Delta\varphi_{sim}$ = simulated phases
$\Delta\varphi_{thermal}$ = thermal noise (normal distribution)
$\Delta\varphi_{coherent}$ = coherent noise (normal distribution)
$\Delta\varphi_{shadow}$ = shadow noise (uniform distribution, [-π,π]).

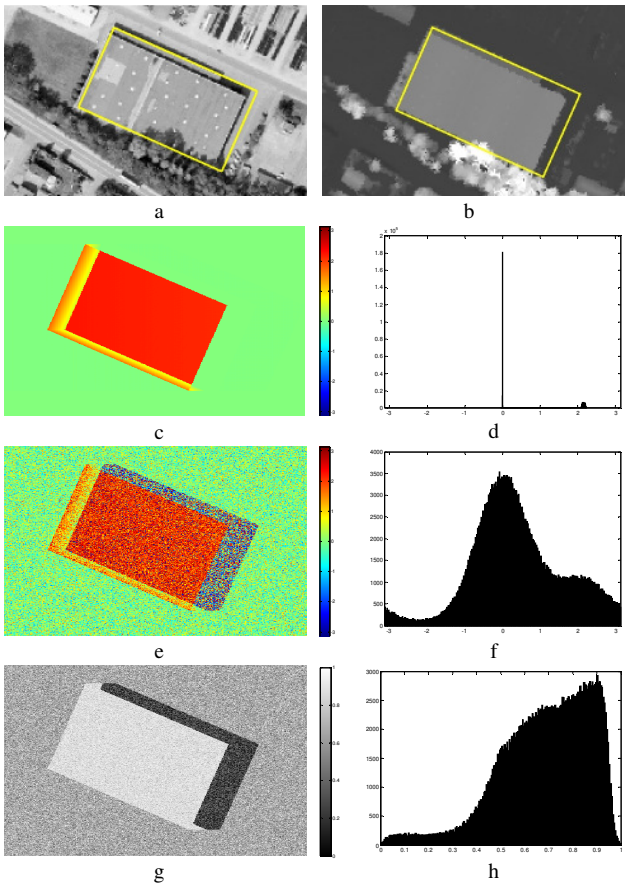


Figure 2. Simulated building signature: a) optical and b) LIDAR image overlaid with GIS information, c) simulated phases and d) histogram, e) noisy simulated phases and f) histogram, g) noisy simulated coherence and h) histogram

The resulting noisy simulated phases are given in Figure 2e with the corresponding histogram in Figure 2f. Additionally, the simulation of a corresponding coherence map was on demand to enable more filter tests on the noisy simulated phases. The

$$coh_{noisy} = coh_{ground} + coh_{building} + coh_{shadow} + coh_{\Delta\varphi} \quad (2)$$

where coh_{ground}, $coh_{building}$, coh_{shadow} are formed by the normal distribution (ndf) of different μ,σ combinations, corresponding to measurements in real InSAR data. The last component $coh_{\Delta\varphi}$ is added to make a connection to noisy simulated phase in pixel space. The resulting noisy coherence map is given in Figure 2g with the corresponding histogram in Figure 2h.

3.2 Real Interferometric SAR Data

The choice of real InSAR data was driven by minimizing noise due to temporal decorrelation, since this kind of noise can be almost neglected for satellite systems like TanDEM-X. Hence, we performed our tests with imagery of the airborne sensor Aes-1 (Schwäbisch et al. 1999). The system operated in X-band at 3000 m flight height with a spatial resolution of about 38 cm in range and 17 cm in azimuth direction. The baseline was about 2.4 m and the scene was illuminated with an off-nadir angle spanning a range from 28° up to 52°.

In Figure 3, the InSAR signature of the industrial building is shown. The direct visual comparison of simulated and real InSAR phase and coherence signatures shows high correlations. Differences in the phase signatures are only visible in the layover area due to occlusion effects caused by closed trees. Coherence values of real InSAR data at the building roof are in some cases lower due to different backscatter properties. Differences in histogram shape are caused by areas less decorrelated (e.g., streets, tree shadows) in the simulated InSAR phases.

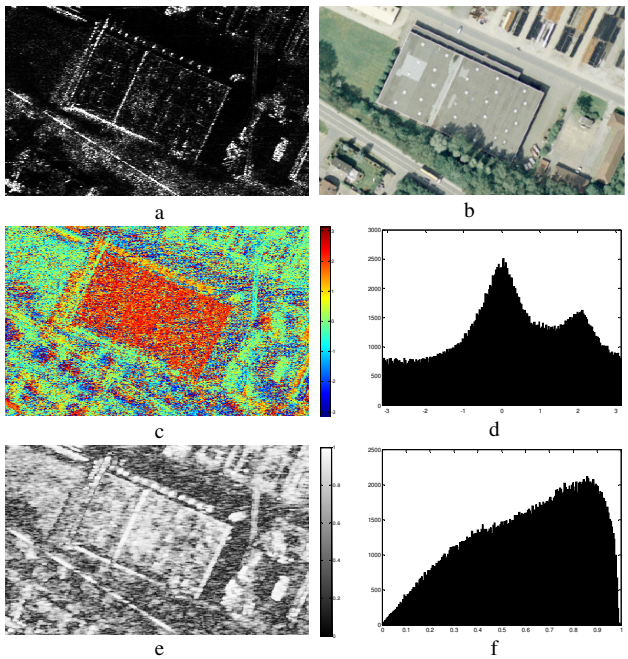


Figure 3. Real building signature: a) SAR magnitude, b) optical image, c) real phases and d) histogram, e) real coherence and f) histogram

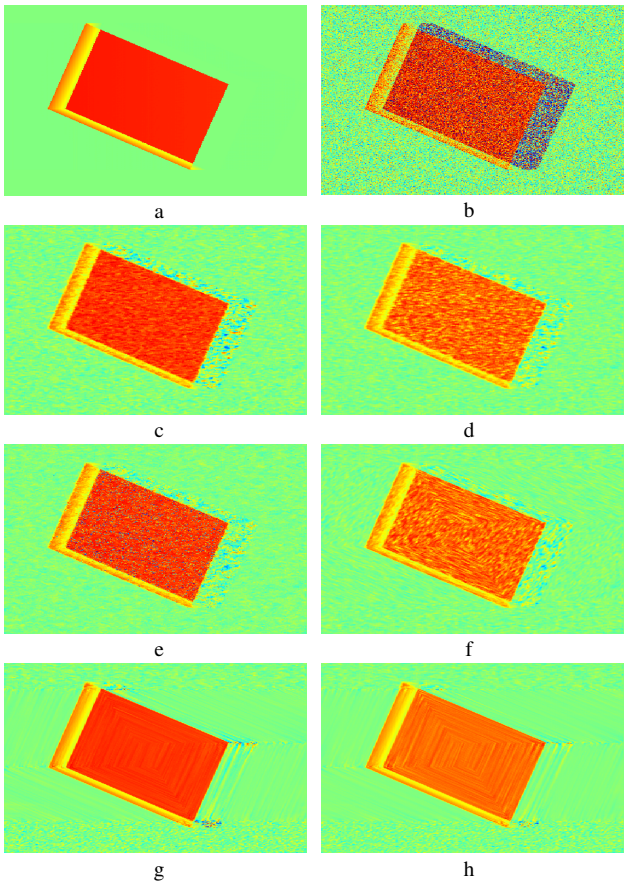


Figure 4. Filtering of simulated interferometric phases: a) original simulation, b) noisy simulation, c) median-filter, d) mean-coherence-filter, e) adapted Lee-filter, using GIS information for f) fixed-window-filter (mean-coherence), g) adaptive-window-filter (median), and h) adaptive-window-filter (mean-coherence)

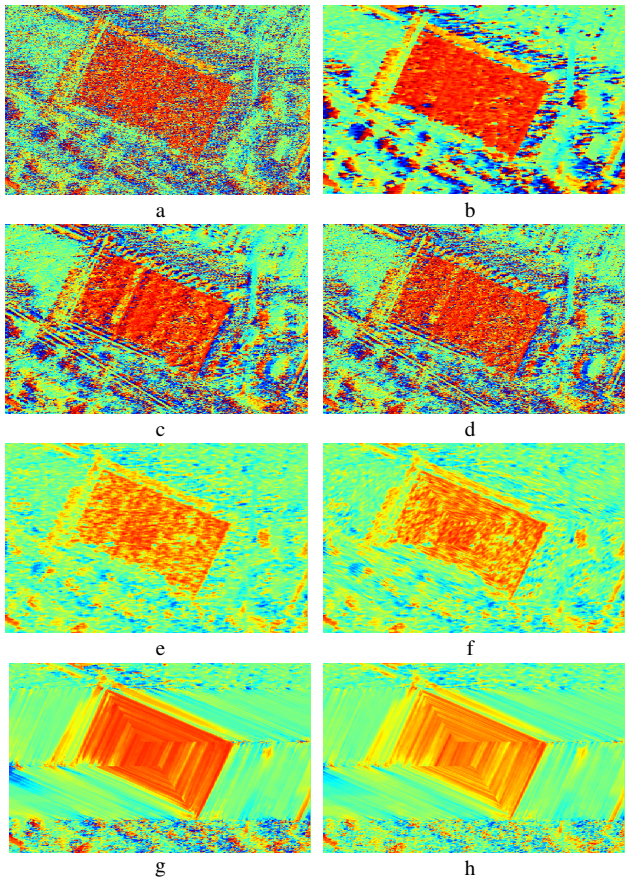


Figure 5. Filtering of real interferometric phases: a) original measured interferometric phases, b) Multilook-filter, c) Goldstein-filter, d) Baran-filter, e) mean-coherence-filter, using GIS information for f) fixed-window-filter (mean-coherence), g) adaptive-window-filter (median), and h) adaptive-window-filter (mean-coherence)

4. VALIDATION OF FILTERING RESULTS

The validation of filter results comprises a visual comparison of filtered building phase signatures and their differences to noise-free simulated phases, as well as a numerical assessment (mean local standard deviation of the filtered phases, variance of phase differences, correlation level between filtered and noise-free simulated phases). For the majority of filters we used a 5 x 5 pixel filter window to achieve comparable results.

4.1 Visual Validation

The filter results on simulated and on real InSAR data are summarized in Figure 4 and 5. The best filter potential concerning phase-based building reconstruction is visible for the new implemented GIS AWFs (Figure 4g,h and Figure 5g,h) as it preserves the shape of the layover area best. Furthermore, tests on the new area filter (mean-coherence, Figure 4d, 5e) and GIS supported FWF show also promising results. Nevertheless, focussing on roof substructures the adaptive filtering can lead to

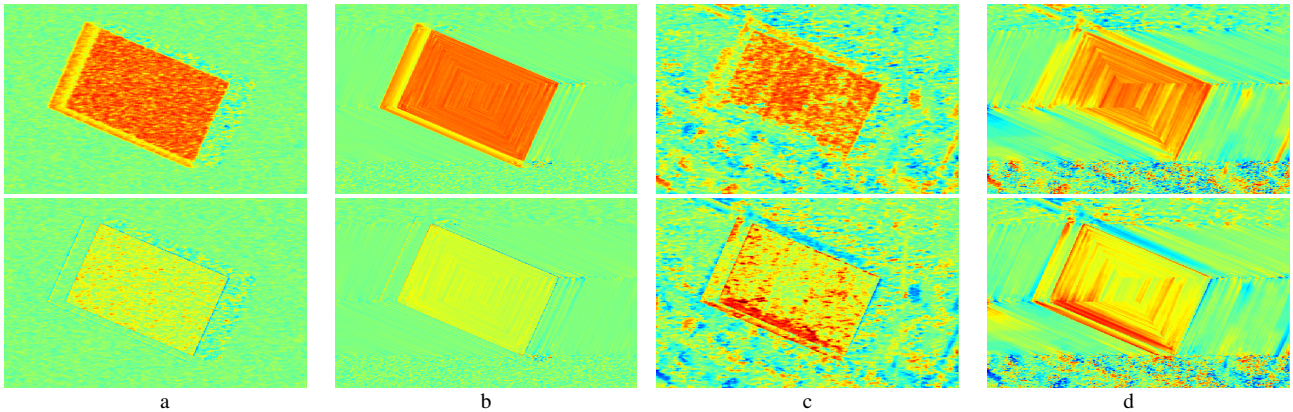
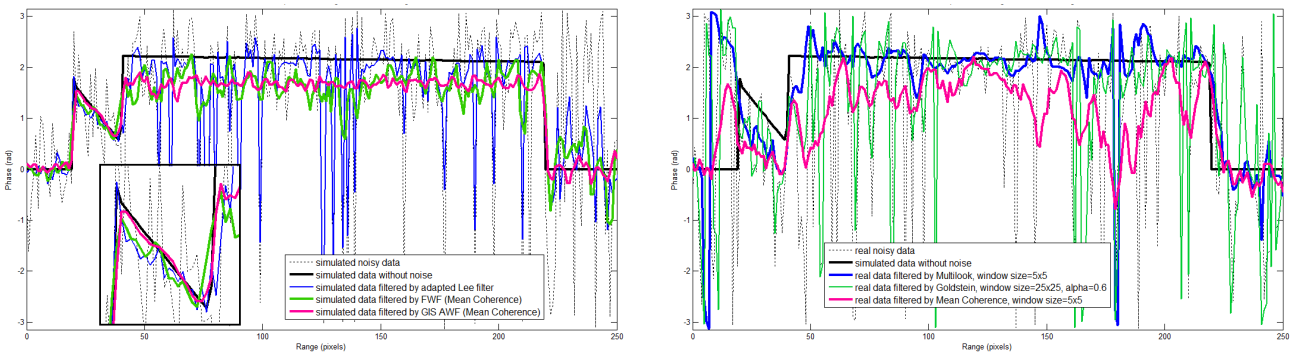


Figure 6. Filter results and differences between of noise-free simulated phases (Figure 4a) and filtered interferometric phases: a) Mean-coherence-filter and b) GIS AWF (mean+coh) on noisy simulated phases, c) Mean-coherence-filter and d) GIS AWF (mean+coh) on real interferometric phases



a b

Figure 7. Slant-Range Profiles: a) filtering of noisy simulated phases and b) filtering of real interferometric phases

blur effects, which can be an advantage or disadvantage depending on application.

In Figure 6, phase plots showing difference between filtered and noise-free simulate phases are presented. The difference maps yield some border effects and a bias at the building roof. The first effect engendered by averaging and chosen filter width is acceptable; the second – probably caused by the missing phase unwrapping – have to be further investigated in future. The differences between real and simulated data (Figure 6c,d) stress the differences between assumed 3D building model and actual situation. In the real data, characteristic layover areas as well as a part of the roof are not visible due to neighboured trees (dark red in the difference map). Furthermore, the considered 3D building model does not contain a building substructure (e.g., canopy) marked blue in the upper part of the building. Hence, the matching of 3D building model and real building essentially affects the filtering result. Additionally, the registration of filter window and real building orientation has to be really good, which depends on the quality of GIS data and illumination / sensor parameters.

A more detailed comparison by focusing on preservation of layover shape and building roof level is possible on the profiles given in Figure 7, whereby a) summarizes the orientation filters and b) the area filters. Best results based on noisy simulated phases are achieved by the GIS AWF (mean+coh). The layover shape is preserved (see zoom-in at Figure 7a), noise is reduced, and only the mean phase level at building roof shows a bias since no local phase unwrapping is considered during filtering. Further investigations are planned at this point.

In the given example, the Multilook-filter shows best results in the group of area filters. The mean roof level is well-preserved and phase unwrapping is not necessary. Hence a combination of the two bests could be improving the filter results once again.

Filter Name	Standard Deviation	Variance	Cross Correlation
noisy simulated phase	0.893	1.015	0.568
Median-filter	0.157	0.062	0.958
Mean-Coherence-filter	0.143	0.097	0.949
Adapted Lee-filter	0.287	0.224	0.850
GIS FWF (mean+coh)	0.155	0.094	0.951
GIS AWF (median)	0.136	0.044	0.971
GIS AWF (mean+coh)	0.120	0.075	0.969

Table 1. Results of filtering based on simulated phases

4.2 Numerical Validation

The numerical validation of our filter results is based on three different values: the mean local standard deviation of filtered phases, the variance of differences (see Figure 6) and the cross correlation between filtered and noise-free simulated phases.
The mean local standard deviation of the resulting interferometric phases $\overline{\sigma_{\Delta\varphi}}$ is defined by:

$$\overline{\sigma_{\Delta\varphi}} = \frac{\sum_{i=1}^{m}\left(\frac{1}{n-1}\sum_{j=1}^{n}(\Delta\varphi_j - \overline{\Delta\varphi})^2\right)^{1/2}}{m} \quad (3)$$

where n = 25, a local window of 5 x 5 pixel and m contains the full image size.
The numerical results are summarized for noisy simulated phases in Table 1 and for real InSAR data in Table 2. Different filters are applied depending on provided data layer. Values given in brackets are only related to building layover and roof area – no ground and shadow area is considered. From the group of area filters shows the modified Median and Mean-Coherence filter best results. An integration of a local phase unwrapping can probably especially improve the results of the Mean-Coherence filter again. For the group of orientation filters with fixed window size (adapted Lee-filter and GIS FWF) an improvement by investigating building orientation from GIS is visible. Results of the adaptive window size filters are similar to the area filter. The AWF Median show slightly better results than the AWF Mean-Coherence implementation. Furthermore, filter results on simulated data show a really high improvement, which is not comparable to tests on real InSAR data. Nevertheless, the high potential on appropriate phase filtering is demonstrated. The benefit of GIS AWFs compared to area filters will turns out more clearly by investigating data showing smaller building, which is planned in future publications.

Filter Name	Standard Deviation	Variance	Cross Correlation
Original phase data	1.365	2.550 (3.595)	0.275 (0.186)
Multilook-filter	0.547	1.352 (0.975)	0.525 (0.496)
Goldstein-filter	1.081	2.182 (2.470)	0.375 (0.307)
Baran-filter	1.1758	2.270 (2.790)	0.349 (0.267)
Mean-Coherence-filter	0.287	0.548 (1.230)	0.628 (0.471)
GIS FWF (mean+coh)	0.291	0.530 (1.204)	0.639 (0.483)
GIS AWF (median)	0.300	0.534 (0.662)	0.681 (0.621)
GIS AWF (mean+coh)	0.241	0.475 (1.007)	0.674 (0.587)

Table 2. Results of filtering based on real InSAR phases

5. CONCLUSION AND OUTLOOK

In this paper, we presented a modified and enhanced filter approaches to smooth interferometric phase signatures of buildings by incorporating available GIS information. Based on the relative pixel position an appropriate window size and window orientation of the phase filter is chosen. First results were shown on simulated and real measured InSAR phases of a single-pass airborne SAR sensor. The visual and numerical interpretation of the new filter result was supported by applying current filter (e.g., Multilook, Goldstein, Lee), too. These first results show a high potential of such GIS-driven filters, especially for applications of building reconstruction.

Our further steps will focus on additional tests considering different airborne and spaceborne data of single-pass and repeat-pass configurations (e.g, TerraSAR-X and TanDEM-X). Furthermore, the abovementioned local unwrapping problem during the phase filtering will be considered. The given solution in Lee et al. 1998 has to be compared with the information given by the phase simulation. Beside, improvements on Goldstein-filter and Baran-filter are planned, as well as the implementation of a combined adaptive GIS and area filter to enable reasonable adjustments to layover and roof areas. Further, investigations on building neighbourhood should be mentioned here, because interaction effects between buildings can necessitate the adaptation of phase filtering.

Finally, the overall goal will be the integration in the superordinate approach of automatic 3D building reconstruction utilizing GIS and mono- or multi-temporal InSAR data to support the following two applications:

- the analysis of mono-temporal TanDEM-X scenes to extract object information for supporting the planning of energy supply in the future, whereby building specific volumes have to be estimated for the evaluation of the energy density. Such data are used to generate building-based thermal maps, which provide another level of planning guides to support the setup of new innovative CO2-reduced heat combined systems.
- the analysis of multi-temporal TanDEM-X scenes. It is targeted to generate automatic comparison between the reconstruction 3D building structures and their new multi-temporal InSAR signature. In case of important changes due to earthquakes or other natural disasters, the identification of changes is an appropriate support for rescuers.

6. REFERENCES

Baran, I., Stewart, M. P., Kampes, B. M., Perski, Z., Lilly P., 2003 A Modification to the Goldstein Radar Interferogram Filter. IEEE Transaction on Geoscience and Remote Sensing, vol. 41, no. 9, pp. 2114-2118.

Bo, G., Dellepiane, S., Beneventano, G., 1999 A locally adaptive approach for interferometric phase noise reduction. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), doi: 10.1109/IGARSS.1999.773466, vol. 1, pp. 264-266.

Bolter, R., 2001 Buildings from SAR: Detection and Reconstruction of Buildings from Multiple View High Resolution Interferometric SAR Data. University Graz: Ph. D. Thesis.

Ciuc, M., Trouve, E., Bolon, P., Buzuloiu, V., 2002 Amplitude-driven coherence filtering in complex interferograms. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), doi: 10.1109/IGARSS.2002.1027213, vol. 6, pp. 3453-3455.

Goldstein, R. M., Werner C. L., 1998 Radar interferogram filtering for geophysical applications. Geophysical Research Letters, vol. 25, no. 21, pp. 4035-4038.

Lee, J.-S., Hoppel, K. W., Mango, S. A., Miller A. R., 1994 Intensity and Phase Statistics of Multilook Polarimetric and Interferometric SAR Imagery. IEEE Transaction on Geoscience and Remote Sensing, vol. 32, no. 5, pp. 1017-1028.

Lee, J.-S., Papathanassiou, K. P., Ainsworth, T. L., Grunes, M. R., Reigber, A., 1998 A new technique for noise filtering of SAR interferometric phase images. IEEE Transaction on Geoscience and Remote Sensing, vol. 36, no. 5, pp. 1456-1465.

López-Martínez, C., Fàbregas, and X., 2002 Modeling and reduction of SAR interferometric phase noise in the wavelet domain. IEEE Transaction on Geoscience and Remote Sensing, vol. 40, no. 12, pp. 2553–2566..

Reeves, B., Homer, J., Stickley, G., Noon, D., Longstaff, I.D., 1999 Spatial vector filtering to reduce noise in interferometric phase images. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), doi: 10.1109/IGARSS.1999.773465, vol. 1, pp. 260-263.

Rejichi, S., Chaabane, F., Tupin, F., Bloch, I., 2010 Morphological filtering of SAR interferometric images. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), doi: 10.1109/IGARSS.2010.5653767, pp. 1581-1584.

Schwaebisch, M., Moreira, J., 1999 The high resolution airborne interferometric SAR AeS-1. Proceedings of the Fourth International Airborne Remote Sensing Conference and Exhibition, Canada, pp. 540-547

Soergel, U., Thoennessen, U., Stilla, U., 2003 Iterative Building Reconstruction in Multi-Aspect InSAR Data. In: Maas HG, Vosselman G, Streilein A (eds) 3-D Reconstruction from Air-borne Laser-scanner and InSAR Data, IntArchPhRS, vol. 34, part 3/W13, pp. 186-192.

Thiele, A., Cadario, E., Schulz, K., Thoennessen, U., Soergel, U., 2007a Building recognition from multi-aspect high resolution InSAR data in urban area. IEEE Transactions on Geoscience and Remote Sensing, vol. 45, no. 11, pp. 3583-3593.

Thiele, A., Cadario, E., Schulz, K., Thoennessen, U., Soergel, U., 2007b InSAR Phase Profiles at Building Locations. Proceeding of ISPRS Photogrammetric Image Analysis, vol. XXXVI, part 3/W49A, pp. 203-208.

Thiele, A., Hinz, S., Cadario, E., 2010 Fusion of InSAR and GIS data for 3D building reconstruction and change detection. Proceedings of 'Fringe 2009 Workshop', Frascati, Italy, ESA SP-677, http://earth.eo.esa.int/workshops/fringe09/.

Tupin, F., 2011 How Advanced Image Processing Helps For SAR Image Restoration and Analysis. IEEE Geoscience and Remote Sensing Newsletter, Cumulative Issue 158, March 2011, ISSN: 0274-6338, pp. 10-17.

Vasile, G., Emmanuel, T., 2004 General adaptive-neighborhood technique for improving synthetic aperture radar interferometric coherence estimation. Journal of the Optical Society of America A, vol. 21, no. 8, pp. 1455-1464.

PHOTOGRAMMETRIC MONITORING OF UNDER WATER EROSION IN THE VICINITY OF CYLINDRICAL BRIDGE PIERS

K. Eder [a], C. Rapp [b], V. Kohl [b], B. Hanrieder [a], U. Stilla [a]

[a] Photogrammetry and Remote Sensing, [b] Fachgebiet Hydromechanik
Technische Universitaet Muenchen, 80333 München, Germany,
konrad.eder@bv.tum.de, rapp@tum.de, stilla@tum.de

Working Group III/5

KEY WORDS: under water photogrammetry, image sequence analysis, close range

ABSTRACT:

Within this work the three dimensional development of a scour around a cylindrical bridge pier has been investigated experimentally. The erosion process has been surveyed in short time intervals using a photogrammetric approach. The results of the work give an unprecedented insight in the development of the scour geometry over time and therefore a crucial step to the fundamental understanding of the phenomena.

1. MOTIVATON AND BACKGROUND

The modelling of sediment transport in turbulent flows is rather complex and up to now there is no general approach available to predict such phenomena. The transport of sediment in water is not only a problem for the navigation of vessels in harbour basins and reservoirs but can be dangerous for the stability of bridge piers (Zanke, 1982).

The scouring process around bridge piers bears immense risks and damages can lead to high expenses. Therefore foundations of bridge piers in rivers are generally uneconomically overdesigned. However, collapses of bridges are not uncommon. In 1990 a highway bridge close to Kufstein, Austria, subsided. The repair estimate was 25 Mio. Euros and the maintenance period lasted more than two years. Fortunately nobody was harmed. During a flood event in Taiwan in 2010 more than one hundred bridges collapsed. One does not have to mention that not only the direct costs but also the impact on the infrastructure and the provision of goods has been a major threat for the people and the industry.

A scour can evolve within flow conditions wherein generally no sediment transport occurs. Scours develop when structures deflect the stream in such a way that the instantaneous velocities and pressures lead to a lift force that acts onto certain grains. These corns are transported and sediment in the wake of the structure. During the scouring process the flow field adjusts to the developing geometry. This fluid structure interaction is a challenging task for the prediction of such processes. However, the so-called horse shoe vortexes that appear at a certain stage stabilize the slopes at the front and at the sides so that a final state is reached (see Figure 1.).

The sediment transport begins as the logarithmic velocity profile of the approaching flow induces a vertical pressure gradient that leads to a down-flow in the vicinity of the pier. This downflow is being deflected and due to the tangential velocity component transported around the cylinder (Pfleger 2011). The turbulence intensity and therefore the lift force decreases in the wake of the pier so that the corns are being deposited.

The investigation of a turbulent flow and the flow-sediment interaction is an interesting field of research that is being investigated by numerous research groups e.g. (Graf and Istiarto, 2002, Malavasi et al., 2004, Melville, 1997 and Oliveto and Hager, 2002). However, the evolution of the scour geometry has been hitherto monitored only in terms of the maximum scour depth. The final state of the geometry was only documented by means of point-by-point laser distance measurements so far (Link, 2006, Pfleger, 2010).

The photogrammetric approach that has been developed within this work gives a first but in time and space resolved insight into the whole scour geometry which leads to a deeper understanding of such phenomena.

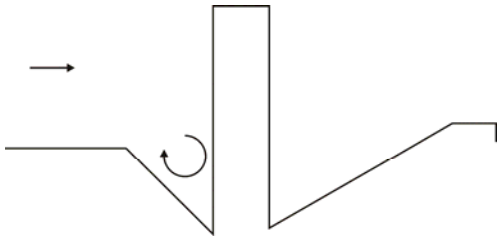


Figure 1. Schematic profile of a cylindrical obstacle in a scour.

2. THE EXPERIMENTAL SETUP

Within a rectangular flume (width 1.20 m) a cylindrical pier (diameter 10 cm) was mounted. The river bed consisted of uniform sand (d_{50} = 1.9 mm) with a mixture of yellow, white and black corns in order to get enough texture for the photogrammetric processing. The sand was levelled to a horizontal plain where a set of 5 ground control points were established by metal sticks with markers on top. The height of the sticks was precisely adapted to the elevation of the initial plain of the river bed. About 15 cm above the river bed a plate of acrylic glass was installed in order to avoid minor waves and to receive a well defined refraction during the experiment (see Figure 2.). On top of this plate another set of 6 ground control points were placed.

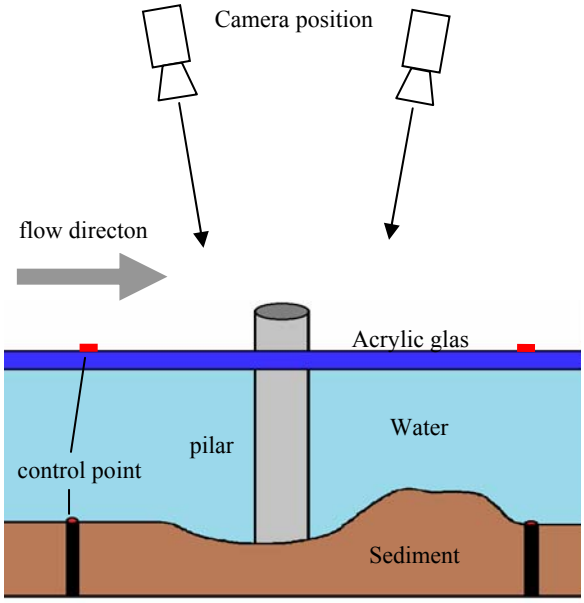


Figure 2. Schematic sketch of the experimental set up

Two digital cameras with a focal length of 24 mm were mounted approximately 2.5 meter above the channel. The used cameras are components of a Particle Image Velocimetry (PIV) system and have a detector array of 4 Megapixels. The camera configuration was calibrated by test field calibration and allows acquiring stereoscopic images taken at the same time. The challenge in this experiment was to investigate the performance of photogrammetric image matching through the plate of acrylic glass and water.

In a first step, the whole setup was imaged without water in order to determine the control points and the riverbed. We call the situation of riverbed the 0-state or "dry state". In the next step, water was flooded to a level just up to the acrylic glass. Again images were taken to study the effect of "under water" image matching. Then the water flow was activated to an extent, where the scour began to form up.

With the synchronized cameras images were taken with a 1 Hz time interval over a period of two hours. During the experiment, about 7200 image pairs were acquired and stored. According to experiences made in previous experiments 42 image pairs have been selected for further processing representing the time resolution as follows:

- Duration of experiment: 0 – 30 minutes: 1 minutes interval
- Duration of experiment: 30 -60 minutes: 5 minutes interval
- Duration of experiment: 60 -120 minutes: 10 minutes interval

3. PHOTOGRAMMETRIC PROCESSING

3.1 Preparation

The photogrammetric processing chain was carried out with the Leica Photogrammetry Suite (LPS), Leica Geosytems, where the parameters especially for image matching had to be optimized for this special application. Precondition for any photogrammetric approach is the reconstruction of the interior and exterior orientation of the camera. The interior orientation parameters for the PIV cameras were determined by test field calibration. The results are given in table1.

Sensor name	PIV_24_left	PIV_24_right
focal length (mm)	24.0180	24.0301
X0 (mm)	-0.0356	0.0654
Y0 (mm)	0.0190	0.0291
rad. distortion (A1)	-2.2183E-004	-2.5384E-004
(A2)	1.2675E-005	1.0594E-005

Table 1. Parameters of interior orientation obtained from testfield calibration

The exterior orientation was reconstructed using the ground control points established on top of the acrylic glass plate, to ensure that there is no displacement of the image ray by acrylic glass and water. The exterior orientation was considered to be constant for the complete image sequence since the camera position did not change. For image matching a proper texture is necessary. A pre-experimental work (Hanrieder, 2010) has shown that a mixture of sand with 10% white and 20% dark particles supplies an optimal texture for image matching. Figure 3 shows a section around the pillar with the mixture of the sediment.

The selected image pairs where then imported into the LPS and assigned with the exterior orientation from the 0-state image pair. For DEM extraction different strategy parameters were tested. The best result was obtained with a modified parameter set of "rolling hills".

Figure 3. The pillar and sediment mixture at 0-state

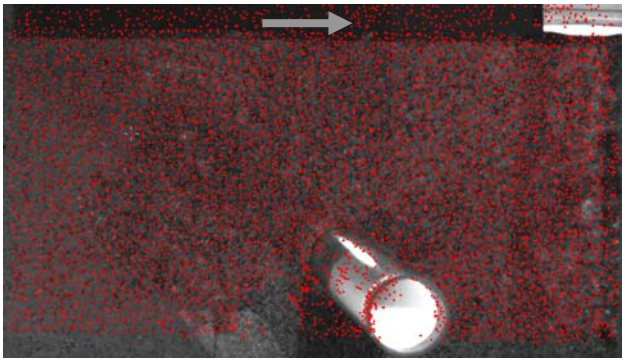


Figure 4. Image of the left camera superimposed with matched points

The image pairs were than processed in batch mode returning about 8000 well distributed object points (Figure 4).

The pier itself caused some occlusions and miss matchings in its near surroundings. These areas were cropped in the final DTM in order to avoid errors in the volume calculation.
From this point cloud a regular grid (spacing 10 mm) was generated. These DEM provided the base for further volume calculations to study the sediment transportation. A visualisation of the final DEM (state 119) is given in figure 5.

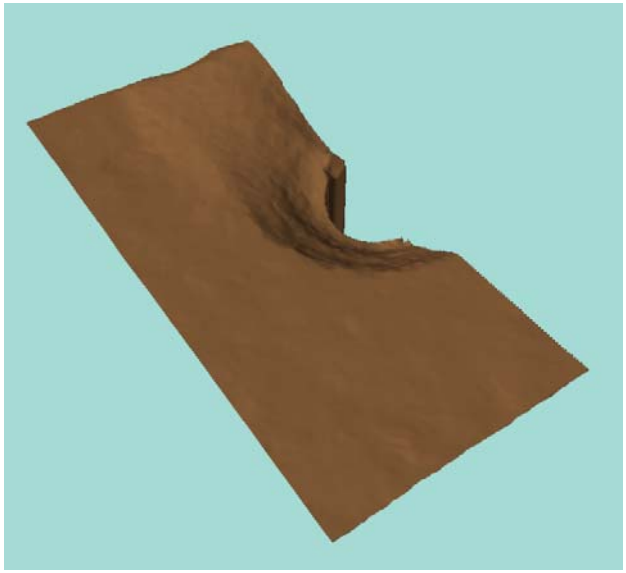


Figure 5. Perspective view of the DEM (state 119 sec.)

3.2 Three media photogrammetry

As stated before, the special challenge of this project was to consider the fact, that three media are involved. Tracing the ray from the camera position to the object point it first crosses the acrylic glass plate and then enters into the water (Figure 6). The refraction indices of these media are:

- Air 1.0
- Acrylic glass 1.49
- Water 1.33

The effect of different media in close range photogrammetry has been a topic of research during the last decades (Höhle 1971; Wrobel, 1975; Li et al., 1997; Butler et al., 2002; Maas, 2008). Also the under water calibration of cameras was investigated (Fryer and Fraser, 1986)

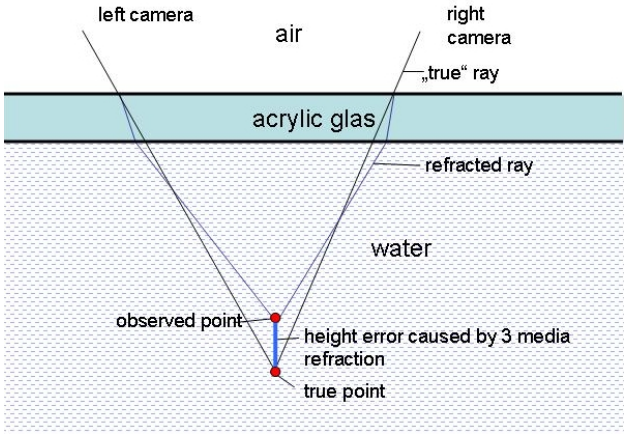


Figure 6. Three media refraction and point displacement

The set up of the control points within this experiment allows for the computation of the differences between “observed“ and “true” points. Since the control points can be measured through air/acryl/water media and their real position is known from the installation, the differences can be calculated (Table 2).

The differences show that the displacement in height is much larger than in planimetry. To compensate the error in this experiment, a rather simple but effective method was applied: The point clouds were transformed by a 7 parameter transformation (shift, rotation, scale), using the five stick points as identical points for the calculation of the transformation parameters. Table 2 shows in the last three columns the residuals at the control points after transformation. This transformation compensates the error at the river bed level, but cannot serve as a rigorous model for the correction of the complete scene.

Point Nr.	Δx [cm]	Δy [cm]	Δz [cm]	res. x [cm]	res. y [cm]	res. z [cm]
900	0,37	0,82	6,33	0.12	0.08	0.06
901	0,57	0,48	6,71	-0.06	0.02	0.02
902	-0,25	0,46	6,92	-0.04	0.01	0.02
903	-0,36	0,89	6,97	0.06	-0.01	0.02
904	0,16	0,99	6,59	-0.07	-0.10	-0.08

Table 2. Differences at control points before and after 3D Transformation

4. HYDROMECHANICAL ANALYSIS

The experiment was conducted at 80% of the critical velocity where sediment transport takes place in an undisturbed flow. The DEM give a clear insight in the evolution process of a scour around a cylindrical bridge pier. This three dimensional geometry information over time has hitherto not been published.

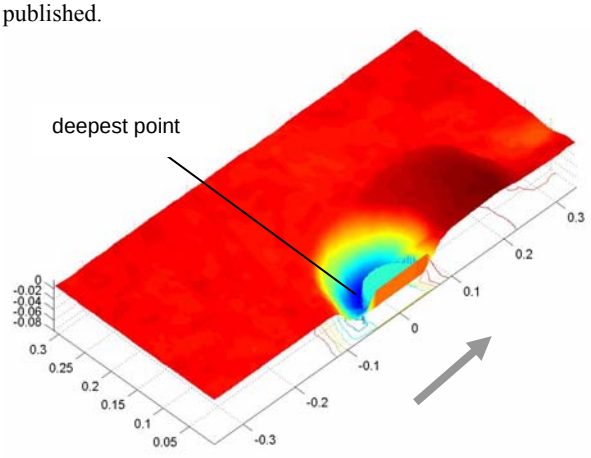


Figure 7. DEM after 17 Minutes. Flow is from lower left to upper right.

Figure 7 shows the DEM-after 17 min of the scouring process and differences between state 1019 and state 0. One can see the cut of sediment around the cylinder with its centre at (x,y) = (0,0). The grains of sand are being transported by the flow and deposited in the wake of the bridge pier. The scour evolves approximately circularly around the pier and extends to about 5 cm around the cylinder. From the figure one can see that the deepest point is not at the pier front.

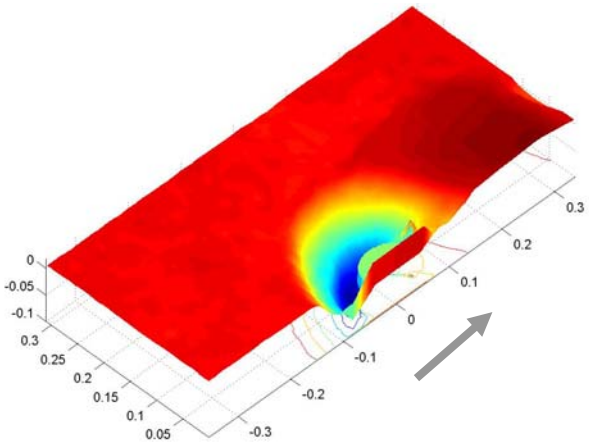


Figure 8. DEM after one hour. Flow is from lower left to upper right.

Figure 8 shows the elevation model one hour after the beginning of the scouring process. The hole has become slightly deeper and wider. The width of the deposition hill corresponds to the width of the scour hole. The end of the hill is approaching the limit of the observation area. In this period of the scouring process the horseshoe vortex system is already stabilizing the geometry. The so called horseshoe vortex system is a characteristic flow structure that evolves around a pier. The current is being deflected downwards by the pier and at the bottom upstream (see Fig. 1) inducing a vortex. This vortex is being transported around the pier before it separates (=horseshoe) and drifts downstream in the wake of the cylinder (Unger and Hager, 2007). Its upward facing component holds the particles at their places so that the slope is steeper than the angle of repose.

The plot shows a growing scour that extends to about 10 cm to the pier front. The deepest point travelled to the pier front and a cut of material can clearly be seen in the wake of the hill. The area of deposition has also increased in time. Its width again corresponds to the width of the scour however its length has increased by about 50% in comparison to the former state. Interestingly the hill foot has been shifted downstream by at least 5 cm.

The transformed and corrected DEM served as input for volume calculation for studying the sediment transportation, especially cut and fill for the selected states representing the sediment movement. One can expect that cut and fill are in the same order such that the mass balance should be nearly zero. The result of the volume calculation is represented graphically in Figure 9.

It shows a positive mass balance over the duration of the experiment. At the beginning (state 0 – state 119) there is a rather large discrepancy. However, Figures 7 and 8 show a clear positive deposition in the remote areas, which can have its origin in general transport in the undisturbed flow.

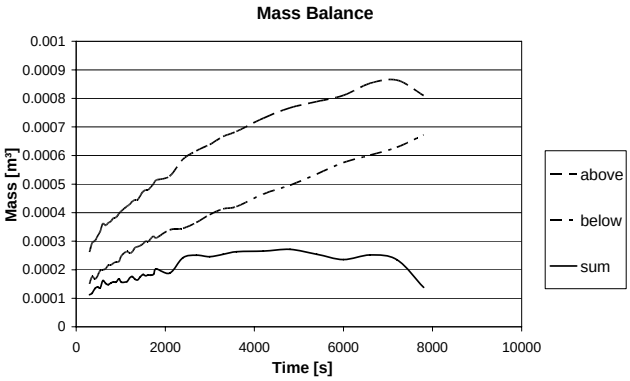


Figure 9. Cut/fill and mass balance during scour development

This effect explains the offset of the mass balance that is applied right at the beginning. However, the mass balance reaches a maximum after one hour and decreases to the end of the experiment. In between 2500 s and 7000 s the mass gain/loss diminishes nearly as expected. The mass decreases after about 7000 s as the more and more particles are transported out of the observation area (see above). However, the mass gain cannot be explained physically, so that further investigations have to be carried out.

5. CONCLUSION AND OUTLOOK

A first insight has been given into the three dimensional time dependent evolution of a scour hole around a cylindrical bridge pier. The highly resolved DEMs can help to explain the complex flow structures that are evolving in such situations. However, the data has to be validated more thoroughly in order to fulfil the requirement of a conservative mass balance. A first approach would take the impact of ray distortion on the elevation model into account. A correction algorithm that uses control points at different levels will be applied.

For continuing former investigations of a single camera setup (Pfleger, 2010) a particle tracking algorithm can be applied to monitor the movement of the grains within these image sequences. First results of the vector field are promising as shown in Figure 10.

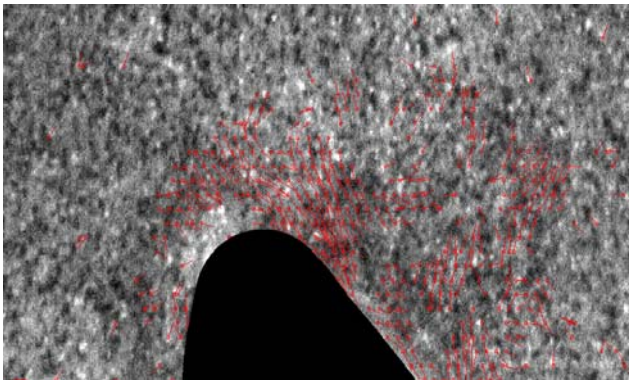


Figure 10. Visualization of tracked particles

REFERENCES

Butler, J.B.; Lane, S.N.; Chandler, J.H.; Porfiri, E. (2002) Through-water close range digital photogrammetry in flume and field environments. Photogrammetric Record, 17 (72): 419-439

Fryer, J.; Fraser, C. (1986) On the calibration of underwater cameras. Photogrammetric Record, 12 (67): 73-85

Graf, W.H.; Istiarto, I. (2002) Flow pattern in scour hole around a cylinder. Journal of Hydraulic Research 40 (1): 13-20

Hanrieder, B. (2010) Photogrammetrische Vermessung von Sedimentoberflächen für Strömungsversuche, Bachelor's Thesis. München: Technische Universtät München, Photogrammetry and Remote Sensing

Höhle, J. (1971) Zur Theorie und Praxis der Unterwasser-Photogrammetrie. Schriften der DGK, Reihe C, Heft 163.

Li, R.; Li, H.; Zou W.; Smith R. G.; Curran, T. A (1997) Quantitative photogrammetric analysis of digital underwater video imagery. IEEE Journal of Oceanic Engineering, 22(2): 364-375

Link, O.; Pfleger, F.; Zanke, U. (2006) Automatic non-intrusive measurement of scour-hole geometry. In: Verheij, H., Hoffmans, G. (eds) Proceedings to the Third International Conference on Scour and Erosion. The Netherlands, Gouda: CURNET, 403-

Maas, H.-G. (1992) Digitale Photogrammetrie in der dreidimensonaöen Strömungsmesstechnik. Dissertation Nr. 9665, Zürich: ETH

Maas, H.-G. (2008) New developments in multimedia photogrammetry. Institute of Geodesy and Photogrammetry, Swiss Federal Institute of Technology

Malavasi, S.; Radice, A.; Ballio, F. (2004) Study of sediment motion in a local scour hole through an image processing technique. In: Proceedings to River flow 2004, II Int. Conf. on Fluvial Hydraulics; 535-542

Melville, B. W. (1997) Pier and abutment scour: integrated approach. In: Journal of Hydraulic Engineering 123 (2):125-136

Oliveto, G.; Hager, W. H. (2002) Temporal evolution of clear-water pier and abutment scour. Journal of Hydraulic Engineering 128 (9): 811-820

Pfleger, F. (2011) Experimentelle Untersuchung der Auskolkung um einen zylindrischen Brückenpfeiler. Dissertation. München: Technische Universität München, Fakultät für Bauingenieur- und Vermessungswesen

Pfleger, F.; Rapp, Ch.; Manhart, M. (2010) Analysis of the temporal evolution of the sediment movement in the vicinity of a cylindrical bridge pier. In: Burns, S. E.; Bhatia, S. K.; Avila, C. M. C.; Hunt, B. E. (eds) Proceedings to the Fifth International Conference on Scour and Erosion, Geotechnical Special Publication No.210, 658-667

Unger, J.; Hager, W.H. (2007) Down-ow and horseshoe vortex characteristics of sediment embedded bridge piers. In: Exp Fluids 42; 1-19

Wrobel, B. (1975) Mehrmedien-Photogrammetrie – ein aktuelles Betätigungsfeld der Photgrammetrie. Vermessungs-wesen und Raumordnung, 37(1)

Zanke, U. C. E. (1982) Grundlagen der Sedimentbewegung. Berlin: Springer

CALIBRATION EVALUATION AND CALIBRATION STABILITY MONITORING OF FRINGE PROJECTION BASED 3D SCANNERS

C. Bräuer-Burchardt, A. Breitbarth, C. Munkelt, M. Heinze, P. Kühmstedt, G. Notni

Fraunhofer IOF Jena, Albert-Einstein-Str. 7, D-07745 Jena, Germany
christian.braeuer-burchardt@iof.fraunhofer.de

Commission III, WG III/1

KEY WORDS: Fringe Projection, Calibration Evaluation, Epipolar Geometry, Optical 3D Measurement

ABSTRACT:

In this work a simple new method for calibration evaluation and calibration stability monitoring of fringe projection based 3D scanners is introduced. This method is based on high precision point correspondence finding of fringe projection sensors using phase values in two perpendicular directions and epipolar geometry concerning calibration data of stereo sensors. The calibration evaluation method can be applied in the measurement process and does not require any additional effort or equipment. It allows the evaluation of the current set of calibration parameters and consideration of the stability of the current calibration over certain temporal progression. Additionally, the quality of distortion correction can be scored. The behavior of three fringe projection based 3D stereo scanner types was analyzed by experimental measurements. Results of the different types of scanners show that calibration may be stable over a long time period. On the other hand, suddenly occurring disturbances may be detected well. Additionally, the calibration error usually shows a significant drift in the warm-up phase until the operating temperature is achieved.

1. INTRODUCTION

Contactless scanning of the surface of different measuring objects is increasingly required in industry and technique, scientific research, cultural heritage preservation and documentation, and in medicine. Fringe projection is the basic principle of a family of such 3D scanners. Flexibility, measuring accuracy, measurement data volume, and fields of application always increase. However, it should be guaranteed that the promised accuracy which is established theoretically can be also achieved under real measurement conditions.

High precision measuring systems based on image data require high precision optical components. The measuring accuracy, however, depends additionally on the quality of the geometric description of the components. The procedure of determination of the geometry of the optical components of a 3D scanning system is performed in the process of camera calibration. The correctness of calibration is crucial for the quality of a photogrammetric system and essential for its measuring accuracy.

The set of calibration data will be produced in the process of initial calibration in the production process of a certain 3D sensor device. However, it is usually not known how stable the initial calibration is over a longer time period. As Luhmann describes (Luhmann et al. 2006), modern photogrammetric measurement systems based on active structured light projection can achieve a measurement accuracy of up to 1:100000 compared to the length extension of the measuring field. However, such high accuracies can only be achieved, if the geometry of the measuring system is stable over the time between calibration and measurement. This is, unfortunately, only the case, if certain measurement conditions strictly hold. However, in the practical use this cannot always be ensured as e.g. reported by Hastedt (Hastedt et al. 2002) or Rieke-Zapp (Rieke-Zapp et al. 2009).

The stability of the calibration data of certain cameras which are used for measurement tasks has been recently analysed by several authors. Mitishita (Mitishita et al. 2003) analyses the interior orientation parameters from small format digital cameras using so called on-the-job-calibration. Shortis (Shortis et al. 2001) examines the calibration stability for a certain digital still camera. Läbe (Läbe and Förstner 2004) determines the geometric stability of low-cost digital consumer cameras. Habib (Habib et al. 2005) analyses the stability of SLR cameras over a long period (half a year). An interesting approach presents Gonzales (Gonzales et al. 2005) in his work. He analyses the stability of camera calibration depending on the proposed calibration technique.

An extensive review of the uncertainty of the epipolar geometry is given by Zhang (Zhang 1998). In his work several techniques for estimating the fundamental matrix and its uncertainty are presented. Dang (Dang et al. 2009) introduces a method for continuous stereo self-calibration by camera parameter tracking based on three different geometric constraints. His work also includes a detailed description of the sensitivity of the 3D reconstruction depending on erroneous calibration parameters.

As it could be observed by consideration of 3D sensor measurements, sometimes systematic errors occur. These errors are characterized by incorrect scaling or deformation of the shape. However, such errors are very difficult to detect, because they are first typically small, and second, the true exact size and the detailed shape of a measuring object is usually unknown. In this work, a novel methodology for calibration parameter evaluation and calibration stability supervision especially for fringe projection based stereo scanners is introduced. Application examples are given for the time dependent behaviour of the calibration quality of different fringe projection based 3D sensors as well as simulation examples.

2. BASIC PRINCIPLES

2.1 Phasogrammetry

Phasogrammetry is the mathematical connection of photogrammetry and fringe projection. The classical approach of fringe projection is described e.g. by Schreiber (Schreiber and Notni, 2000). It can be briefly outlined as follows. A fringe projection unit projects well defined fringe sequences for phase calculation onto the object, which is observed by a camera. Measurement values are the phase values obtained by the analysis of the observed fringe pattern sequence at the image coordinates [*x, y*] of the camera. The 3D coordinates *X, Y, Z* of the measurement point *M* are calculated by triangulation, see e.g. (Luhmann et al., 2006). The calculated 3D coordinate depends linearly on the phase value.

2.2 Epipolar Geometry

The epipolar geometry is a well-known principle which is often used in photogrammetry when stereo systems are present. See for example (Luhmann et al., 2006). It is characterized by an arrangement of two cameras observing almost the same object scene. A measuring object point *M* defines together with the projection centres O_1 and O_2 of the cameras a plane *E* in the 3D space (see also Figure 1). The images of *E* are corresponding epipolar lines concerning *M*. When the image point *p* of *M* is selected in camera image I_1 the corresponding point *q* in camera image I_2 must lie on the corresponding epipolar line. This restricts the search area in the task of finding corresponding points. In the following we consider a system consisting of two cameras C_1 and C_2 and one projector in a fix arrangement.

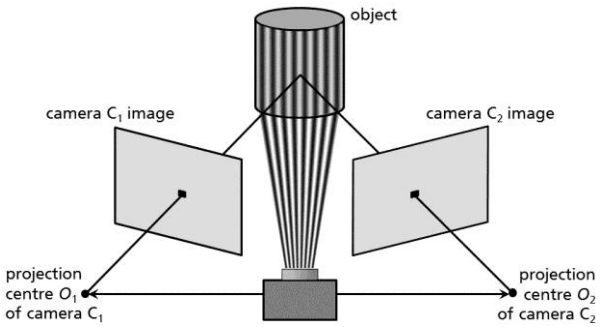


Figure 1: Stereo camera arrangement with fringe projector

2.3 Camera Calibration

Camera calibration is the determination of the intrinsic and extrinsic parameters (including lens distortion parameters) of an optical system. It has been extensively described in the literature, e.g. by Luhmann (Luhmann et al. 2006), Chen (Chen and Brown 2000), Brown (Brown 1971), Tsai (Tsai 1986), or Weng (Weng et al. 1992). Different principles are applied to perform camera calibration. The selection depends on the kind of the optical system, the exterior conditions, the effort to be pushed, and the desired measurement quality. In case of the calibration of photogrammetric stereo camera pairs, the intrinsic parameters of both cameras should be determined as well as the relative orientation between the cameras. Intrinsic parameters include principal length *c*, principal point $pp = (u, v)$, and distortion description.

The position of the camera in the 3D coordinate system is described by the position of the projection centre O = (X, Y, Z) and a rotation matrix R obtained from the three orientation angles ω (pitch angle), ϕ (yaw angle), and κ (roll angle). Considering stereo camera systems, only the relative orientation between the two cameras (Luhmann et al. 2006) has to be considered, because the absolute position of the stereo sensor is usually out of interest. Lens distortion may be considerable and should be corrected by a distortion correction operator D. Distortion may be described by distortion functions as e.g. proposed by Tsai (Tsai 1986), or Weng (Weng et al. 1992), or by a field of selected distortion vectors (distortion matrix), as suggested by Hastedt (Hastedt et al. 2002) or Bräuer-Burchardt (Bräuer-Burchardt et al. 2004).

3. APPROACH FOR CALIBRATION EVALUATION

3.1 Basic Assumptions

Let us consider a fringe projection based stereo scanner with two cameras and one projector in a certain arrangement (see also Figure 1). 3D surface data are aquired by a measurement from one sensor position using either epipolar constraint (mode m_1) or fringe pattern sequences in two orthogonal projection directions (mode m_2 - see Schreiber and Notni, 2000). Point correspondences with sub-pixel accuracy are found using equal phase values in both camera images. Whereas using mode m_2 always guarantees to find correct point correspondences, searching on epipolar lines (mode m_1) leads to point correspondence errors. However, mode m_1 needs half the length of the image sequence and hence image sequence recording needs half the time only.

The calibration of the sensor is performed once in the laboratory and is assumed to be fix over a longer time period. Calibration data include intrinsic and extrinsic parameters and a set of distortion parameters or a distortion matrix for each camera. Calibration may be performed again, of course, if it seems to be necessary. However, this means usually high effort. Alternatively, self-calibration (Schreiber and Notni 2000) can be performed at every measurement. However, conditions may be too poor to obtain robust parameters from self-calibration.

3.2 Goal

Especially moved sensors are susceptible to mechanic shocks and vibrations. These influences may disturb the current calibration, and calibration parameters become erroneous. Additionally, it is well known that calibration data are temperature dependent. Thermic changes in the environment or increase of working temperature of the sensor may also influence the calibration parameters.

The goal of this work is to describe the current state of the set of calibration parameters of fringe projection based stereo sensors. Small errors can be compensated by parameter correction whereas considerable errors should imply the decision to perform a new calibration.

3.3 Epipolar Line Error

Obviously the correct position of the epipolar lines strongly depends on the accurate values of the calibration data set. Therefore calibration errors directly influence the correctness of point correspondences, if searched on the epipolar lines. Hence, if corresponding points are located exactly on the corresponding epipolar lines this implies that the calibration data are “good”.

However, this is not sufficient for the statement that the calibration data are correct. Certain calibration parameters may be disturbed without having a considerable influence on the position of the epipolar lines but leading to a disparity error which leads, subsequently, to a depth error of the reconstructed 3D points. On the other hand, if the epipolar line position is erroneous, calibration data must be disturbed.

Let us consider two corresponding points p_i and q_i. The epipolar line position error $err_{pos}(p,q)_i$ is defined as the perpendicular distance of the correct corresponding point q_i to the epipolar line g_i, defined by p_i and the set of calibration parameters. Further the *rms epipolar line error* ΔE_{rms} of the image pair is defined by

$$\Delta E_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(err_{pos}(p,q)_i\right)^2} \qquad (1)$$

where n is the number of considered corresponding point pairs. These points should be well distributed over the images. Additionally, the maximal epipolar line error ΔE_{max} can be considered: $\Delta E_{max} = \max\{\ |\ err_{pos}(p,q)_i\ |\}$, $i = 1, \ldots, n$. Alternatively, a value of the 95% percentile may be chosen for ΔE_{rms} in order to prevent using outliers. Actually, epipolar line position error $err_{pos}(p,q)_i$ is signed in contrast to *rms epipolar line error* ΔE_{rms}. This is meaningful, because the sum of all $err_{pos}(p,q)_i$ may be near zero whereas the amount of *rms epipolar line error* ΔE_{rms} may be considerable.

3.4 Analysis of Calibration Parameter Error Influence

Let us consider the parameters of interior orientation c_i, u_i, v_i, exterior orientation X_i, Y_i, Z_i, ϕ_i, ω_i, κ_i, i=1, 2 (altogether 18 parameters) and the distortion D_i, describing the geometry of the stereo sensor. Because of symmetry reasons (only relative orientation between the cameras is of interest) only twelve parameters (omitting X_1, Y_1, Z_1, ϕ_1, ω_1, κ_1) are considered. Note that the meaning of the parameters may be different regarding either terrestrial or aerial orientation, respectively.

Assume an aerial orientation of the stereo camera pair with small pitch (or tilt) angles ω_i, a yaw (or gear) angle difference $\Delta\phi$ between the cameras near triangulation angle, and small roll (or rotation) angles κ_i for both cameras. Assume X-axis alignment approximately parallel to the baseline. Estimation of the influence of erroneous calibration parameters and image co-ordinates to the epipolar line error err_{pos} was performed. It was derived from intercept theorems and collinearity equations (see e.g. Luhmann et al. 2006). Results of error influence estimation are given in table 1, where Δx and Δy are the centered image coordinates, and d is the distance to the measuring object. A more detailed analysis of the error influence was performed by Dang (Dang et al. 2009).

3.5 Calibration Evaluation

The idea for calibration evaluation is to simply describe the calibration quality by the amount of the *rms epipolar line error* according to equation (1). In order to get a significant value, representative points should be extracted being well distributed over the images. Hence, a virtual grid of image coordinates is defined. The number n of grid points should be at least $n = 100$. For all points p_i in the image of camera C_1 the corresponding points q_i in the image of camera C_2 are found by use of the rotated phase algorithm.

Parameter error	Influence on err_{pos}
ΔX	$err_{pos} \approx 0$
ΔY	$err_{pos} \approx \Delta Y \cdot c/d$
ΔZ	$err_{pos} \approx \Delta y \cdot \Delta Z/c$
$\Delta\phi$	$err_{pos} \approx 0$
$\Delta\omega$	$err_{pos} \approx c \cdot \tan(\Delta\omega)$
$\Delta\kappa$	$err_{pos} \approx \Delta x \cdot \tan(\Delta\kappa)$
Δc	$err_{pos} \approx \Delta y \cdot \Delta c/c$
Δu	$err_{pos} \approx 0$
Δv	$err_{pos} \approx \Delta v$

Table 1: Influence of calibration parameter errors for a certain aerial arrangement

3.6 Approach for Calibration Correction

The knowledge of the amount of the mean epipolar line error implies the idea to manipulate the calibration parameters such that the mean epipolar line error becomes minimal. This was performed with the help of a newly developed algorithm, which is described by the authors (Bräuer-Burchardt et al. 2011).

It can be briefly explained as follows. In a first step an analysis of the influence of the single calibration parameters is performed. A reduced set of parameters with considerable and different sensitivity (see next section) is selected. The remaining parameters (in our experiments between three and seven) are systematically changed thus that a minimization of the mean epipolar line error is achieved. This will be achieved by an iterative algorithm. The manipulated parameters are now used as the current calibration data. Usually, using this method the *rms epipolar line error* may be reduced to a value of near or below 0.1 (pixel). This is very accurate and allows finding point correspondences on the epipolar lines with high precision.

This algorithm may be extended performing the minimization of the scaling error, too. Scaling error realizes the consideration of those parameters having a poor sensitivity to the epipolar line error (Bräuer-Burchardt et al. 2011). However, for the determination of the scaling error additional information is necessary which cannot be extracted from the phase data of an arbitrary measuring object. Here an object with well-defined lengths should be used, e.g. a grid pattern.

4. EXPERIMENTS AND RESULTS

4.1 Simulation of Calibration Parameter Error Influence

First, simulations showing the possible difference of the influence of the single parameters on the epipolar line error were performed. It must be noted that the influence of the parameters strongly depends on the actual geometric arrangement of the sensor. Hence, simulations are performed in such a way, that the geometric situation of a real sensor is approximated, i.e. a sensor in aerial arrangement with a measuring field of about 20 mm x 15 mm and a triangulation angle of about 15°. A meaningful error of each of the parameters of the exterior and interior orientation of the sensor was assumed. The same parameters of the two cameras show analogous influence. Some parameters show similar behaviour as others ($Y \sim \omega \sim v$, $X \sim \phi \sim u$). The results confirm the theoretic analysis (see table 1).

4.2 Simulation Example of a 3D Measurement

In order to illustrate the effect of an erroneous calibration parameter set simulations was performed showing a scene with four spheres with two different diameters (1 mm and 2 mm) and a frustum of a cone, respectively. 3D reconstruction was performed using original and disturbed data. The (meaningful) amount of disturbance (ΔE_{rms} = 1, 2, and 5, respectively) was chosen according to experimental results. See Figures 2 and 3 which illustrate the effect of disturbed calibration parameters. Considering the spheres, it can be seen that the effect of parameter errors is hardly detectable from the subjective evaluation of the sphere shape, even if the error is big (five pixel). However, decreasing completeness can be observed. Contrary, the results of the frustum of a cone imply that errors can be detected earlier by user observation.

Table 2 documents the error ΔD of the sphere diameter measurement and the maximal length measurement error Δlen_{mx} obtained by the distance measurement between two sphere centre points. Additionally, the mean standard deviation SD_{mn} of the 3D points on the sphere surfaces is given. The size of the simulated scene was about 20 mm x 15 mm, the sphere diameters D were D_1 = 1.98 mm and D_2 = 3.96 mm and the distance between the spheres 11.9 and 16.8 mm, respectively. It can be seen that epipolar line error has a considerable and non-linear influence on the diameter error. Moreover, the smaller the diameter the bigger is the percentage error. Length error, however, is weak and increases proportionally.

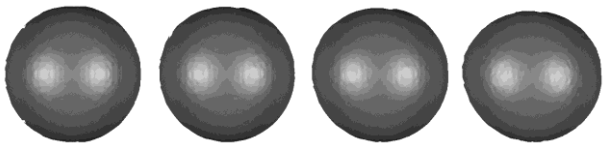


Figure 2: Sphere measurement (simulation) with ΔE = 0, ΔE = 1, ΔE = 2, and ΔE =5 (from left to right)

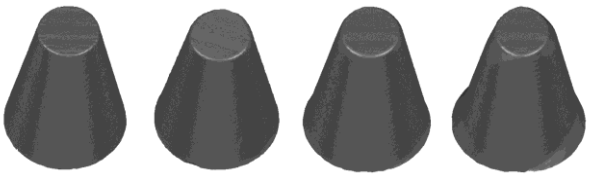


Figure 3: Simulated frustum of a cone measurement with ΔE = 0, ΔE = 1, ΔE = 2, and ΔE = 5 (from left to right)

ΔE [pixel]	ΔD_1 [µm]	ΔD_1 [%]	ΔD_2 [µm]	ΔD_2 [%]	Δlen_{mx} [µm]	Δlen_{mx} [%]	SD_{mn} [µm]
0	-	-	-	.	-	-	2,6
1	-5,4	0,27	-1,6	0,04	-2,5	0,015	2,9
2	-18	0,90	-3	0,08	-5	0,03	3,1
5	-70	3,50	-22	0,54	-12	0,07	5,0

Table 2: Influence of calibration parameter error ΔE on measurement errors of two spheres

4.3 Experiments on Real Data

In the following a number of experiments using real sensor data are documented. The temporal course of the epipolar line error is considered using three different sensor types (see Figure 4). The first sensor DS (see Kühmstedt et al. 2007) has a measurement field of about 20 mm x 14 mm. The image size is 516 x 778 pixel leading to a local resolution of about 35 µm. Two sensors DS_1 and DS_2 were analysed. The second sensor HS (see Bräuer-Burchardt et al. 2011) has a measurement field of about 50 mm x 40 mm. The image size is 2448 x 2048 pixel leading to a local resolution of 20 µm. The third sensor CS (see Munkelt et al. 2007) has a measurement field of about 240 mm x 175 mm. The image size is 640 x 474 pixel leading to a local resolution of 350 µm. Two devices CS_1 and CS_2 of sensor type CS were analysed.

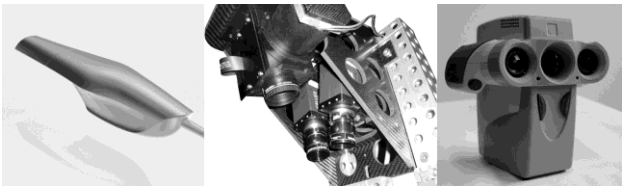


Figure 4: Scanning devices DS (schematic image - left), HS (middle), and CS (right)

Sensors DS and CS are hand held, light weight and mobile scanners with low spatial resolution whereas sensor HS is a high resolution sensor (5 Mpixel) designed for flat measuring objects. More details are given by the authors (Kühmstedt et al. 2007, Munkelt et al. 2007, Bräuer-Burchardt et al. 2011).

The measurements were performed as follows. A white nearly plane object was chosen as measurement object. Measurements of the coordinates of corresponding points of rectangular grid points well distributed over the image plane were performed using a 16 phase algorithm of orthogonally rotated phase directions. It can be assumed that the point correspondence error is negligible. For each corresponding pair of points the epipolar line position error was determined by the perpendicular distance of the measured point q_i to the calculated position of the epipolar line g_i. Mean epipolar line error was calculated using equation (1) and used as feature size of the calibration quality.

By the measurements two aspects of calibration stability should be checked. First, the stability over a longer time period (several weeks) was analysed. For this experiment evaluation measurements were selected with warmed up sensors. The result of epipolar line error of the sensor DS_1 is illustrated in Figure 5.

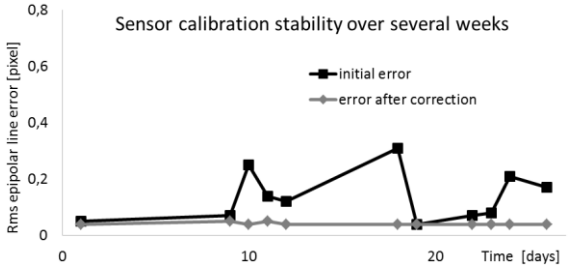


Figure 5: *Rms epipolar line error* and corrected *rms epipolar line error* of scanner DS_1 over a period of 26 days

It can be seen that the calibration remains stable over the quite long period of almost four weeks. At the days of "higher" errors, operating temperature may be outside the target range. However, the operating temperature should be in a certain well-defined range. The next experiment was the consideration of the

calibration stability from switching on the device until the reaching of the operating temperature. These experiments were performed for all three scanners and repeated several times.

By Figures 6 and 7 the behaviour of epipolar line error over 120 minutes of the *rms epipolar line error* of scanner DS is documented for two days. Decreasing error in the "warm up" phase is due to the heating up of the scanner to operating temperature of about 40°. The warming up time takes between 30 and 45 minutes.

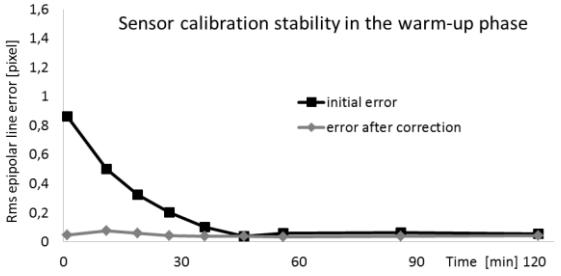


Figure 6: *Rms epipolar line error* of scanner DS_1 over two hours from switch on, 2011/02/07, measuring object: plane

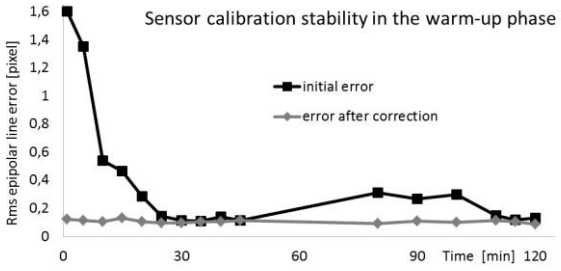


Figure 7: *Rms epipolar line error* of scanner DS_2 over two hours from switch on, 2011/05/13, measuring object: teeth arc

Sensor HS was first analysed over three weeks. See results documented in Figure 8. It can be seen that the calibration is very stable until day 20. Between day 20 and day 21 something happened which disturbed the calibration. However, correction was always possible with a remaining rms epipolar line error below 0.1 pixel.

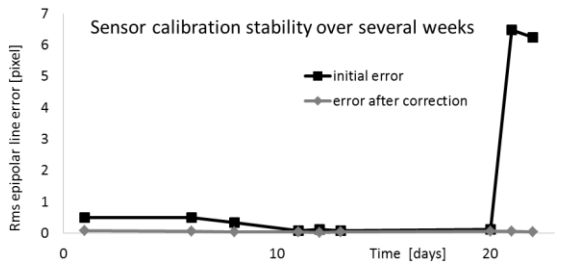


Figure 8: *Rms epipolar line error* and corrected *rms epipolar line error* of scanner HS over a period of 22 days

Next, the behaviour in the warm-up phase was checked at five days. See the results of one measurement illustrated by Figure 9. The progress of the *rms epipolar line error* at the other days was almost identic.

Last but not least sensors of type CS were checked. Figure 10 shows the *rms epipolar line error* over 30 minutes from switch on of sensor CS_1. Scanner CS_2 showed similar behaviour.

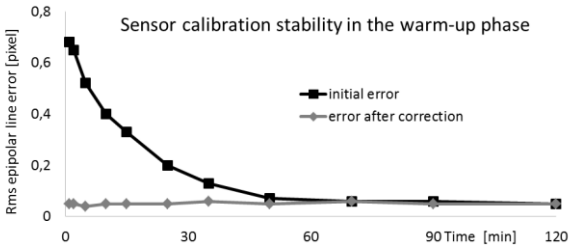


Figure 9: *Rms epipolar line error* of scanner HS over two hours from switch on, 2011/04/06, measuring object: plane

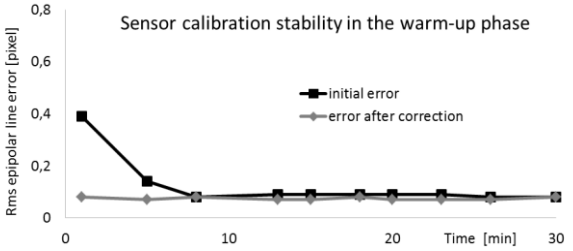


Figure 10: *Rms epipolar line error* of scanner CS_1 over 30 min. from switch on, 2011/02/28, measuring object: plane

4.4 Detection of Erroneous Distortion Description

Epipolar line error can be also used to detect insufficient distortion correction. If correction of radial distortion is not applied correctly, this implies a certain epipolar line error. See Figure 11 which illustrates the effect of disturbed calibration parameters on the epipolar line error.

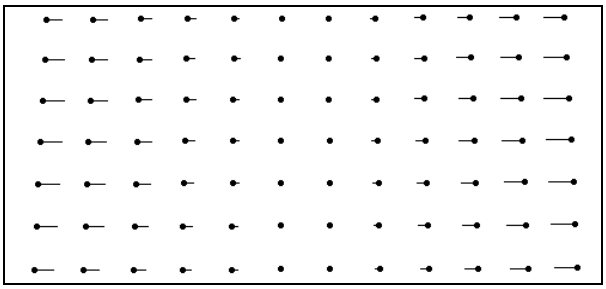


Figure 11: Effect of insufficient distortion correction of CS_2: the vectors indicate epipolar line error scaled by factor ten, epipolar lines run vertically, image section is cutted above and below

5. DISCUSSION AND OUTLOOK

The three analyzed scanner types show similar behavior of the calibration quality represented by the *rms epipolar line error*. After being switched on the sensors need a certain time to reach their operating temperature. The duration of this warm-up phase is different depending on the properties of the sensors and can be exactly determined by the proposed method.

Sudden changes of the calibration data which may be caused by mechanic influences as shocks or vibrations are reliably detected by the method. Depending on the shape of the measuring object these changes are not necessarily noticed by the user as it showed the sphere example. However, subsequent correction may be performed by improvement of the current set of calibration parameters. We proposed a systematic and

iterative change of between three and seven calibration parameters and minimization of the epipolar line error (Bräuer-Burchardt et al. 2011). However, it should be analyzed, whether the proposed correction is sufficient in the current case. Because only the position of epipolar lines is optimized, the corrected set of calibration parameters is only better than the erroneous one, but not really true. New calibration has possibly to be performed in the case of strongly disturbed calibration.

Future work should be addressed on the improvement of the current calibration correction procedure, the monitoring of several 3D scanners over longer periods (several months), and the fully automation of the developed algorithms so far. When scaling error can be determined, a correlation analysis of the *rms epipolar line error* and scaling error should be performed. This will possibly allow a blind correction of more calibration parameters than three.

Furthermore, we plan to develop a new algorithm which realizes the correction of the calibration parameter set using an estimation of the fundamental matrix (see Zhang 1998). First attempts showed a low numeric robustness but the method should be improved.

6. SUMMARY AND CONCLUSION

In this work a simple new method for calibration stability monitoring of fringe projection based 3D scanners was introduced which allows considering the stability of the current calibration over certain temporal progression. The behavior of three types of fringe projection based 3D stereo scanners was analyzed by experimental measurements.

Only in the case of occurrence of calibration parameter errors without showing the epipolar line error effect the proposed method is not sensitive and will not detect a disturbed calibration. However, this case is very unlikely.

Calibration evaluation by the proposed method should be applied if epipolar constraint is used in order to realize point correspondences and the stability of the set of calibration parameters may be disturbed by thermic (warm-up) or mechanic (shocks or vibrations) influences. Correction is necessary according to the amount of the detected error and the requested measuring accuracy. The proposed method is ideal for determination of the end of the warm-up phase and for correction of the point correspondence in the warm-up phase.

REFERENCES

Bräuer-Burchardt, C., 2005. A new methodology for determination and correction of lens distortion in 3D measuring systems using fringe projection. In Pattern Recognition (Proc 27th DAGM), Springer LNCS, pp. 200-207

Bräuer-Burchardt, C., Breitbarth, A., Kühmstedt, P., Schmidt, I., Heinze, M., Notni, G., 2011. Fringe projection based high speed 3D sensor for real-time measurements. Proc. SPIE, Vol. 8082, 808212-1 - 808212-8

Bräuer-Burchardt, C., Kühmstedt, P., Notni, G., 2011. Error compensation by sensor re-calibration in fringe projection based optical 3D stereo scanners. accepted paper at ICIAP 2011, Proc. ICIAP, Ravenna, Springer LNCS, (in print)

Brown, D.C., 1971. Close-range camera calibration. Photogram. Eng. 37(8), 855-66

Chen, F., Brown, G.M., 2000. Overview of three-dimensional shape measurement using optical methods. Opt. Eng. 39, 10-22

Dang, T., Hoffmann, C., Stiller, C., 2009. Continuous stereo self-calibration by camera parameter tracking. IEEE transactions on image processing 18(7), 1536-1550

González, J.I., Gámez, J.C., Artal, C.G. , Cabrera, A.M.N., Stability study of camera calibration methods, 2005. CI Workshop en Agentes Fisicos, Spain, WAF'2005,

Habib, A.F., Pullivelli A.M., and Morgan, M.F., 2005. Quantitative measures for the evaluation of camera stability, Opt. Eng. 44, 033605-1 - 033605-8

Hastedt, H., Luhmann, T., Tecklenburg, W., 2002. Image-variant interior orientation and sensor modelling of high-quality digital cameras. IAPRS 34 (5), 27-32

Kühmstedt, P., Bräuer-Burchardt, C., Munkelt, C., Heinze, M., Palme, M., Schmidt, I., Hintersehr, J., Notni, G., 2007. Intraoral 3D scanner, Proc SPIE Vol. 6762, pp. 67620E-1 - 67620E-9

Läbe, T., Förstner, W., 2004. Geometric Stability of Low-Cost Digital Consumer Cameras. *In Proceedings of the ISPRS Congress*, Istanbul, Turkey, 528-535

Luhmann, T., Robson, S., Kyle, S., Harley, I., 2006. Close range photogrammetry. Wiley Whittles Publishing

Mitishita, E., Côrtes, J., Centeno, J., Machado, A., Martins, M., 2010. Study of stability analysis of the interior orientation parameters from the small format digital camera using on-the-job calibration. Proc ISPRS, XXXVIII, part1

Munkelt, C., Bräuer-Burchardt, C., Kühmstedt, P., Schmidt, I., Notni, G., 2007. Cordless hand-held optical 3D sensor. Proc. SPIE Vol. 6618, pp. 66180D-1

Rieke-Zapp, D., Tecklenburg, W., Peipe, J., Hastedt, H., Haig, C., 2009. Evaluation of the geometric stability and the accuracy potential of digital cameras – Comparing mechanical stabilisation versus parameterisation. ISPRS, Vol. 64/3, 2009, 248-2

Schreiber, W. and Notni, G., 2000. Theory and arrangements of self-calibrating whole-body three-dimensional measurement systems using fringe projection technique. OE 39, 159-169

Shortis, M.R.; Ogleby, C.L.; Robson, S.; Karalis, E.M.; Beyer, H.A., 2001. Calibration modelling and stability testing for the Kodak DC200 series digital still camera. In *Proceedings of SPIEVideometrics and Optical Methods for 3D Shape Measurement*, San Jose, CA, USA, January 2001; pp. 148-153

Tsai, R., 1986. An efficient and accurate camera calibration technique for 3-D machine vision. *IEEE Proc CCVPR*, 364-74

Weng, J., Cohen, P., Herniou, M., 1992. Camera calibration with distortion models and accuracy evaluation. PAMI(14), No 11, 965-80

Zhang, Z., 1998. Determining the epipolar geometry and its uncertainty: a review. IJCV 27(2), 161-198

SIMULATION OF CLOSE-RANGE PHOTOGRAMMETRIC SYSTEMS FOR INDUSTRIAL SURFACE INSPECTION

T. Becker [a], M. Özkul [a], U. Stilla [b]

[a] BMW Group AG, Petuel Ring, 80788 München – {tobias.becker, muammer.oezkul}@bmw.de
[b] Photogrammetry and Remote Sensing, Technische Universität München, 80290 München, Germany - stilla@tum.de

KEY WORDS: Photogrammetry, Simulation, Virtual Reality, POV Ray, Close-Range

ABSTRACT:

Close-range photogrammetric measurement systems are increasingly used for high-precision surface inspection of car body parts. These measurement systems are based on an active light source, the projector, and one or more cameras. Many systems use a sequence of fringe projection, mostly a combination of the gray code and phase shift technique. Basically the quality of the measurement result depends on best possible positions of these sensors, which requires human expert knowledge and experience. But is it possible to use computer-based algorithms to find optimal measuring positions? Simulation processes are discovered as part of a research project aimed at the evaluation of the quality of measuring positions concerning to visibility, the attainable accuracy and realizable feature extraction. One approach is the simulation of the photogrammetric sensor using ray tracing techniques to create photorealistic pictures from the sensor cameras view. This image sequence could be processed with the evaluation software of the system manufacturer in order to calculate a three dimensional point cloud. Following an actual/target comparison should indicate differences that trace back to insufficient measuring positions. In this paper we show how to build up a virtual close range photogrammetric sensor using POV Ray, a free ray tracing software. After introducing the simulation concept, the design of a virtual close range photogrammetric sensor is presented. Based on practical examples of sampled scenes the potential of photorealistic ray tracing is shown. Finally the usability of this simulation approach is discussed.

1. INTRODUCTION

1.1 Motivation

In automotive industry the quality test of car body parts concerning to geometric accuracy and dimensional stability is very important. So far mostly tactile coordinate measuring machines are used for high precision measurement of specified points. Meanwhile there is an increased use of close range photogrammetric systems as they provide industrial compatibility and high accuracy, i.e. Oezkul (2009) researched the practical application of optical measurement systems to detect very small surface defects on body parts. A certain advantage of photogrammetric systems is the contactless and rapid area-based measurement. Figure 1 shows a typical industrial inspection cell where the photogrammetric sensor is mounted on a robot, providing a highly automated process. The practical application has shown that the correct operation of these machines requires expert knowledge and experience, due to a lot of different parameters that affect the measuring process. Major parameters affecting the accuracy are measuring volume, angle between sensor and object, occlusion, shadowing, reflections, contrast/exposure. Furthermore there are some more practical hints like robot configuration or collision. The latest developments of commercial sensor planning tools take care of some of the listed parameters e.g. occlusion, nevertheless they cannot guarantee a sufficient result at single or all measuring positions. These circumstances lead to the fact, that extensive tests within the inspection cell are necessary to ensure best possible results. During these tests the inspection cell is blocked and not usable for the main task – measuring. With a computer-aided simulation tool it should be possible to evaluate the quality of measuring positions concerning to visibility, the attainable accuracy and realizable feature extraction, outside of the inspection cell. Furthermore this evaluation should not base on detailed information in underlying algorithms of individual system manufacturers.

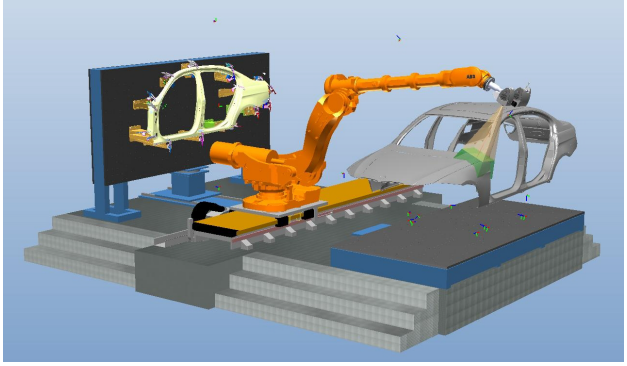


Figure 1. Industrial surface inspection cell

Using virtual reality for simulation of photogrammetric tasks is the subject of a number of publications. The simulation of sensors with the description of objects, sensors and light sources is subject of Raczkowsky and Mittenbuehler (1989) or Ikeuchi and Robert (1991). Determine sensor positions by avoiding occlusion in urban scenes using VRML is described by Stamos and Allen (1998).
Piatti and Lerma (2006) use a virtual simulator for photogrammetry for the determination of proper exposure stations, i.e. in urban areas with high buildings, in terms of getting maximum ground coverage with a minimum of image overlapping.

Auer et al. (2009, 2010) present an approach to simulate SAR images via calculating the elevation data using ray tracing algorithms of POV Ray.
Raguse and Wiggenhagen (2003) simulate the exposure configuration to evaluate the optical measurement chain. Virtual reality is not used in the truest sense only the imaging geometry, i.e. coordinates of the cameras und object points, and some constants are used to make a statistical statement of the network configuration.

1.2 Overview

In this paper we describe a possible simulation concept based on computer generated photorealistic images. Modelling a close range photogrammetric sensor with its camera and projector is one of the main tasks in this concept. In Section 2 the modelling of a fictive sensor and the scene is explained. Section 3 shows some results of rendered images. In section 4 the simulation concept is discussed.

2. SIMULATION CONCEPT

The proposed simulation concept is based on computing photorealistic pictures from the sensors camera view and to process them with the evaluation software of the system manufacturer. There are several tools to compute more or less photorealistic images from three dimensional scenes. For example OpenGL or DirectX visualise three dimensional scenes in real time using hardware acceleration. Both have the disadvantage of missing essential features like shadows or reflections. Other software tools use time intensive ray tracing algorithms to calculate more realistic images.
We use POV Ray (The Persistence of Vision Ray Tracer v3.6), a free and open source ray tracing software tool, to compute images with high photorealistic quality. The main simulation concept consists of the following parts (Figure 2):

- Modelling of the photogrammetric sensor
- Modelling of the scene within the inspection cell
- Calculating the photorealistic image sequence from the sensor view
- Digitize a 3D point cloud from these images
- Carry out an actual/target comparison.

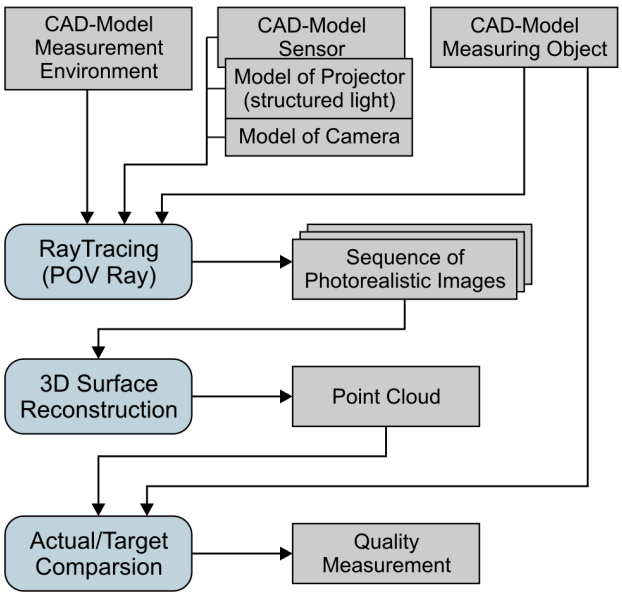


Figure 2. Simulation concept

The major advantage of this concept is that we have not the constraints to know details of the algorithms implemented by the manufactures. In fact they use highly optimized and specialized algorithms for feature extraction like reference points, edges or bolts. These algorithms are based on the captured image, the computed point cloud or a combination of both and are mostly subject to secrecy.
This paper is focused on modelling the different components of photogrammetric sensors and ray tracing.

2.1 Modelling of the photogrammetric sensor

The following section describes the modelling of a photogrammetric sensor with one projector and one or more cameras.
Figure 3 shows a simple model of a photogrammetric sensor with one projector and one camera. In general the sensor has a fixed triangulation angle $\alpha \approx 20°$ between the camera and projector. The measurement volume of the sensor is defined by the depth of focus of the camera and the focus range of the projector. The center of the cameras depth of focus is termed as TCP (tool center point). According to α and the distance from camera to the TCP the basis length b – the distance between projector and camera – is defined. The sensor used in our experiments has a basis length b=350mm corresponding to a triangulation angle α=20° degrees between projector and a mean measuring distance of t=650mm. In the image the visual range of the camera is marked green and the focus range of the projector is marked blue.
For a more detailed description of photogrammetric sensors the reader is referred to Luhmann (2003).

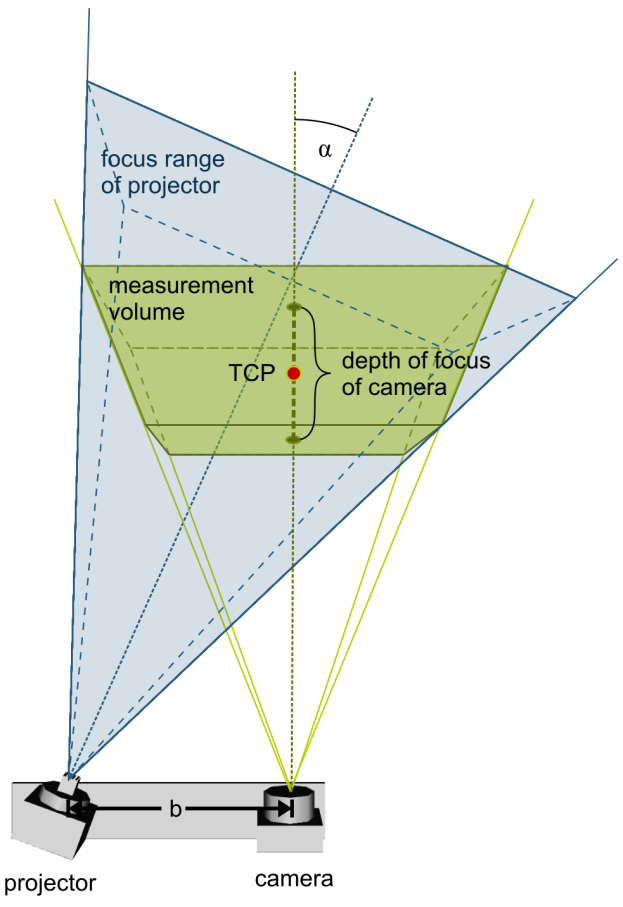


Figure 3. Sensor with measurement volume

2.1.1 Camera: Close range photogrammetric measurement systems mostly use high precision digital cameras with fixed lenses. The manufacturers pay close attention to high grade camera-lens-combination to minimize optical aberrations that would result in distortions of the measurement. The use of fixed lenses leads to a defined depth of focus so that the visual range is a frustum of a pyramid. Parts of objects that appear in front of or behind the frustum of pyramid are not applicable for a high precision measurement and are rejected by the elevation software. POV Ray uses a simple pinhole camera with different projection types, for example perspective, ultra-wide-angle or fisheye. The camera definition describes the position, projection type, line of sight, sky vector, aspect ratio and horizontal viewing angle. Additionally the focal blur can be simulated by specifying a focal point and aperture value. Large apertures give a lot of blurring, while narrow apertures will give a wide zone of sharpness. Since further aberrations of the photogrammetric camera are considered by the calibration tools of the evaluation software of the manufacturers in a pre-processing step, the pinhole camera model of POV Ray in conjunction with focal blur simulation should be sufficient.

2.1.2 Projector: The photogrammetric measurement systems that are commonly assembled in the automotive industries mostly use a combination of the gray code and phase shift technique. The gray code is used to identify the number of the corresponding strip of the phase shift. The projector himself is integrated in a case and consists of a high intensity light source emitting a beam through a glass sheet, with dark etched or dark coloured stripes, and a lens towards the measuring part. So the stripe patterns of the gray code are being displayed as sharp shadows on the surface of the measuring object. The modelling of the projector in POV Ray is achieved by a point light source which is shielded by a case on five sides. In view direction the case is open to integrate a geometry that simulates the stripe pattern for each image in the sequence. The stripe pattern is simulated by polygons integrated in the open side of the case. For each image of the gray code sequence the number of stripes, there dimensions and positions are calculated.
The phase shift technique uses sinusoidally intensity modulated fringes that are projected on the measuring object. In a number of steps the fringes are shifted and for each phase position an image is recorded. The phase shift allows precise calculation of the coordinates with subpixel accuracy. In POV Ray textures are utilized to simulate sinusoidally intensity modulated fringes. For that a texture bitmap has to be set up with a 32-bit RGBA color space, where the RGB values are set to zero, defining a black image. The alpha channel is set to the values of the calculated modulation, defining the fringes via transparency. Now the projector case in the scene is closed by a polygon which is overlaid with the texture. For each phase shift a suitable texture bitmap is calculated.
In most cases the photogrammetric measurement systems uses two additional images, one with full light of the projector but without any stripes, and another image with light turned off. These two images are used to determine the maximum edge contrast of the stripes on the measuring object. With POV Ray it is just the same, using the original projector model without simulated stripes once with and once without light switched on.

2.2 Modelling of the scene

Modelling a complete scene is done with the POV Ray scene description language. A plain ASCII text file is used to describe the scene in a readable and convenient way. A large amount of different geometric objects, effects and global settings are available, also mathematic expressions or macros are possible. As the modelling of the scene may become very complex, we developed a software tool, which contains a framework for visualisation via OpenGL, a scene graph for various geometric objects, parameter input and control for the sensor modelling and more. This tool allows creating 3D scenes or parameterised sensor models in a simple way. An export feature creates automatically the input files for POV Ray of the complete image sequence including all necessary parameter settings.

3. EXPERIMENTS AND RESULTS

3.1 Experiment

With the above mentioned tool the measuring object, a front wring, and a model of a fictive photogrammetric sensor is imported and positioned, so that a portion of the wing lies within the visual range of the sensors camera (see Figure 4).

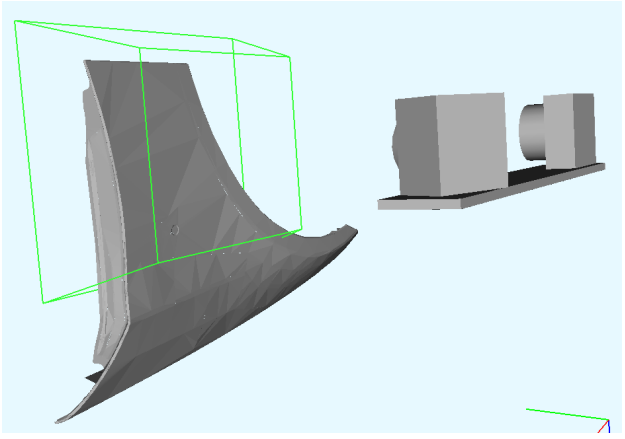


Figure 4. Modelling of the scene

The fictive photogrammetric sensor uses a sequence of 14 images, one with light switched off, one with light switched on but no stripes, eight images with the 8-bit gray code and four images for the phase shifts, each shift by one fourth phase. For a better visualisation of the sinusoidally intensity modulated fringes we use broad stripes. The resolution of the sensor is set to a width of 1600 and height of 1200 pixels. During the export absolute coordinates of the positioned objects are calculated and textures or materials are assigned. For each image in the sequence a separate text file is created. Finally POV Ray is used for sampling each scene.

3.2 Results

Figures 5-7 show the simulated sensor images, first the one without any stripes (light turned on), then a projected gray code and last a phase shift.
An enlarged view of the sinusoidally intensity modulated fringes is shown in Figure 8.

Figure 5. Render results without stripes

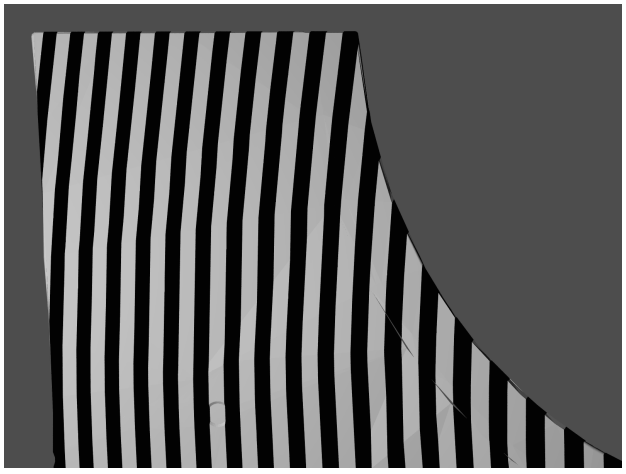


Figure 6. Render result with gray code (7th-Bit)

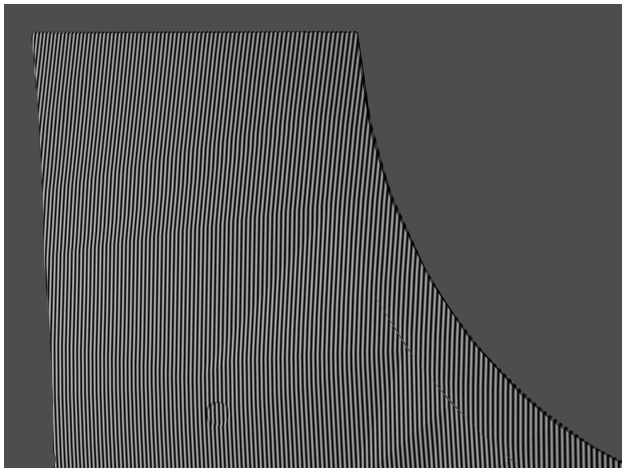


Figure 7. Render result with phase shift

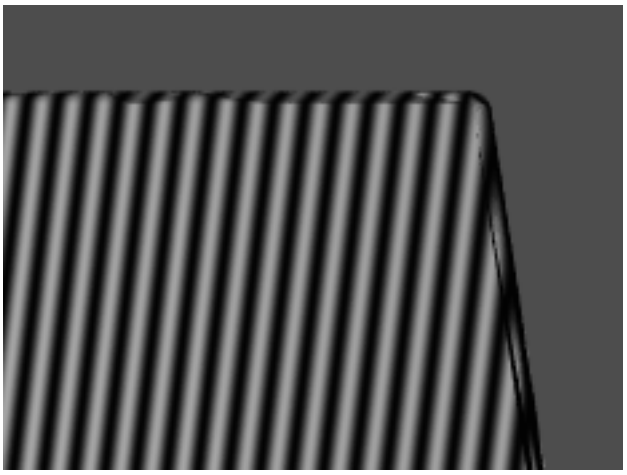


Figure 8. Enlarged view of phase shift

4. DISCUSSION

The simulated image sequence shows a geometric concordance to recorded images of real photogrammetric sensors, especially for the fringe projection of the gray code. One problem is the texture of the measuring object, which is much more complex in reality, probably caused by impurities like small oil stains or partial reflections.

A main drawback is the time needed to sample the images. With an average PC each sampling of the scene took about 5-8 minutes induced by focal blur and antialiasing parameters. This leads to calculation times of nearly two hours for a complete image sequence of one measuring position. It must be said that large body parts need about 50 or more measuring positions, leading to an overall calculation time of about 4 days and more for a complete simulation of the measuring process. However there are approaches to allow realtime ray tracing using highly efficient hardware architecture for example by Schmittler et al. (2002) or Wald (2004).

5. FUTURE WORK

Based on the parameters of a real photogrammetric sensor it should be possible to make a comparison between real and rendered images. Furthermore the computed photorealistic images are used to calculate a point cloud of the measuring object by the software of the measurement system manufacturer. An actual/target comparison should show the geometrical accuracy of the simulation concept.

In addition the modelling of the camera could be enhanced to simulate some realistic chromatic aberration and also some distortion by adding a lens shaped object just in front of the camera.

ACKNOWLEDGEMENTS

This research is funded by the BFS (Bayerische Forschungsstiftung) within the contract number AZ-876-09.

REFERENCES

Auer, S.; Zhu X.; Hinz S.; Bamler R., 2009. Ray Tracing and SAR-Tomography for 3D Analysis of Microwave Scattering at Man-Made Objects. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science*, Vol. 38 (3/W4), pp. 157-162.

Auer, S.; Hinz, S.; Bamler, R., 2010. Ray Tracing Simulation Techniques for Understanding High Resolution SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 48, pp. 1445-1456.

Ikeuchi, K.; Robert, J., 1991. Modeling Sensor Detectability with VANTAGE Geometric/Sonsor Modeler. *IEEE Transactions on Robotics and Automation*. Vol. 7, pp. 771-784.

Luhmann, T., 2003. *Nahbereichsphotogrammetrie Grundlagen, Methoden und Anwendungen*. 2. Auflage. Herbert Wichmann Verlag, Heidelberg, Germany.

Oezkul, M., 2009. *Qualitätsansprüche bezüglich des äußeren Erscheinungsbildes von Automobilen der Premiumklasse*. Hieronymus Verlag, Munich, Germany.

Piatti, E. J.; Lerma, J. L., 2006. A Virtual Simulator For Photogrammetry. *ISPRS Commission V Symposium "Image Engineering and Vision Metrology"*, Vol. 36 (5).

Raczkowsky, J.; Mittenbuehler, K. H., 1989. Simulation of Cameras in Robot Applications. *Computer Graphics Applications*, pp. 16-25.

Raguse, K.; Wiggenhagen, M., 2003. Beurteilung der Optischen Messkette durch Simulation der Aufnahmekonfiguration. In: *Seyfert, E. (Hrsg.): 23. Wissenschaftliche Jahrestagung der DGPF. Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation*, Vol. 12, pp. 275-283.

Schmittler, J.; Wald, I.; Slusallek, P., 2002. SaarCOR – A Hardware Architecture for Ray Tracing. In: *Proceedings of the ACM SIGGRAPH/Eurographics Conference on Graphics Hardware*, pp. 27-36.

Stamos, I.; Allen, P., 1998. Interactive Sensor Planning. In: *IEEE Conference on Computer Vision and Pattern Recognition, Proceedings*, pp. 489-494.

Wald, I., 2004. *Realtime Ray Tracing and Interactive Global Illumination*. Dissertation, Computer Graphics Group, Saarland University, Germany. http://www.sci.utah.edu/~wald/PhD/wald_phd.pdf (accessed 15 May 2011)

DEM GENERATION BY MEANS OF NEW DIGITAL AERIAL CAMERAS

J. Höhle

Dept. of Planning, Aalborg University, Fibigerstraede 11, DK-9220 Aalborg, Denmark, jh@land.aau.dk

Commission I, WG 2

KEY WORDS: Acquisition, Sensor, CCD, Imagery, Impact Analysis, Accuracy, DEM/DTM

ABSTRACT:

Based on practical experiences with the first generation of digital cameras the impact of four new digital aerial cameras on the DEM generation is estimated. The cameras are evaluated with respect to elevation accuracy, area coverage and image quality. In the derived formulae for the absolute accuracy of elevations a camera factor is applied, which summarizes several features in one number. The produced graphs show the potential of the new cameras regarding the relative and absolute accuracy of automatically derived elevations. For better absolute elevation accuracy the increase of the base/height ratio and of the resolving power of lenses are more important than the reduction of the pixel size. Improvement of the elevation accuracy by means of multi-ray photogrammetry is discussed.

1. INTRODUCTION

10 years use of digital aerial cameras has nearly removed film-based aerial cameras. The production of orthoimages and DEMs became completely digital, to a large extent automatic and with short response time. These are the major reasons for the success. Regarding the obtainable elevation accuracy with the first generation of the large format frame cameras it was proven in recent tests that they did not perform much better than analogue cameras (Haala et al., 2010). The format is in reality only a fraction of the format of the analogue camera and the base/height ratio is much lower. Several charge-couple devices (CCD) have to be stitched together in order to obtain a larger format. The pixel size of the applied CCD could be reduced in new versions of digital cameras. Small pixel sizes require efficient electronics and storage devices to read out the huge amount of data in a short time. Also lenses with a higher resolving power are necessary to match the small pixel in order to take advantage of high resolution CCDs. Improvement of the economy in aerial photography has been another wish from the camera users. Three camera manufacturers have announced new large format frame cameras, which represent a second generation. Their potential with regard to the elevation accuracy is to the interest of the mapping industry. The development in aerial photogrammetric cameras of one manufacturer can be read from Figure 1. The depicted three cameras represent three generation of cameras: The analogue camera (RMK Top15), the first generation of digital large format frame camera (DMC) and the latest design (DMCII-250). The image area of digital cameras is considerably smaller than the 23cm x 23cm format of film-based aerial cameras. The field of view (FOV) of the selected lenses is also narrower. The panchromatic image of the DMCII-250 camera is produced by a single lens and a single CCD only. Its 250 Megapixel CCD of DALSA has a very small pixel of 5.6µm x 5.6µm (Intergraph, 2010). In order to obtain imagery of the same ground sampling distance (GSD) the flying altitude for the DMCII-250 has to be two times higher than the altitude when using the DMC. The area covered by one image on the ground is, however, 2.4 times larger than the area covered by the DMC. Beside the mentioned camera three other cameras became recently available: The UC-Xp WA and UC Eagle of Microsoft (Microsoft, 2011) and IGI-235 of IGI (IGI, 2010). Their potential regarding DEM generation will be discussed. It is the objective of this contribution to analyze the influencing factors and to predict the results regarding the elevation accuracy at DEM generation with the four new cameras.

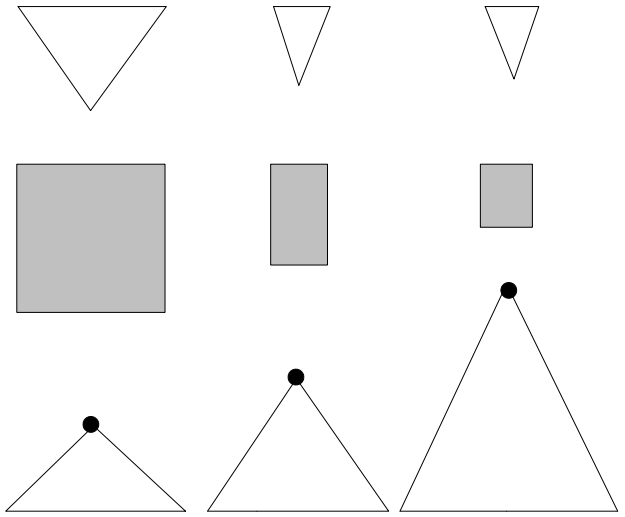


Figure 1. Characteristics of aerial cameras of Intergraph. The symbols mean: From left to right: Film-based wide angle camera, digital cameras DMC and DMCII-250. From top to bottom: FOV in the direction of flight, format, flying height for the same GSD, and FOV across the direction of flight.

2. INNOVATIONS

The innovations in the four mentioned cameras first of all concern the number of pixels per image. It reaches now up to 260 million pixels (Mpixel) in the latest design (UC Eagle). The side of the squared pixel is as small as 5.2µm. The most recent camera of Intergraph uses one large CCD only which avoids the fusion of several images into one image of a larger format. The image size in flight direction of the DMCII-250 has now 14656 pixels corresponding to 82.07mm. It is, however, only 36% of

the film-based camera. The term "large frame" is related to the number of pixels and not to the physical size of the frame. The Table 1 contains all four new cameras with their features which have influence on the geometric accuracy.

The camera factor (K_{cam}) combines several features in one number. It is derived by:

$$K_{cam} = \frac{c}{b' \cdot pel'} \quad (1)$$

where c=camera constant
b'=image base at forward overlap of p=60%
pel'=pixel size.

Features	IGI 235	UC-Xp WA	UC Eagle	DMCII 250
pixel size [µm]	6	6	5.2	5.6
camera constant [mm]	80	70	80	112
image size (in flight direction) [pel] [mm]	12 750 76.50	11 310 67.86	13 080 68.02	14 656 82.07
GSD at flying height of h=1000m [cm]	7.5	8.6	6.5	5.0
base/height ratio at p=60%	0.38	0.39	0.34	0.29
K-factor [μm^{-1}]	0.436	0.430	0.565	0.609

Table 1. Features of four new digital aerial cameras

If the images have the standard forward overlap of p=60%, the image base is $b'=0.4 \cdot s'_{fld}$, where s'_{fld}=image size in flight direction.

From Table 1 it can be noticed that the camera factor (K_{cam}) is lowest for the IGI-235 camera and highest for the DMC-250 camera. The older digital cameras have even smaller factors: 0.412 (UC-D) and 0.269 (DMC). High elevation accuracy is achieved with low camera factors. Other camera features like frame rate and image quality have also influence on the generation of DEM and will also be dealt with in the following.

2.1 Geometric accuracy

The performance of digital aerial cameras has recently been investigated by several research groups (Jacobsen, 2011), (Haala et al., 2010), (Spreckels et al., 2010), (Höhle, 2009a&b). Besides the camera also the flight parameters (altitude, side lap) and the processing tools influence the accuracy of derived elevations. Furthermore, the terrain type, the density, and the definition of the points have an effect. The assessment of the accuracy needs checkpoints of superior accuracy. The applied accuracy measures should not be affected by blunders.

In the investigation of the DMCII-140, which uses **one** 140 Mpixel CCD with a 7.2µm x 7.2µm large pixel, a vertical accuracy of $RMSE_h=0.7 \cdot GSD$ has been achieved for well defined checkpoints using semi-automated measurements (Jacobsen, 2011). The imagery used 65% forward overlap and 65% side overlap and the determination of the check- and control points could in average use seven images. The object points have been determined within an aerotriangulation. The achieved high accuracy is, therefore, not a result of a DEM generation.

In tests of a project group of the German society of photogrammetry, remote sensing and geoinformation (DGPF) it was found that the generation of very dense DEMs by means of the DMC images with GSDs of 8cm and 20cm could be done with $RMSE_h=0.4 \cdot GSD$ and $RMSE_h=0.8 \cdot GSD$ respectively (Haala et al., 2010). The checkpoints were well defined and with good contrast to the surroundings, and the derived DEM had a very narrow spacing of the grid points. A few blunders were eliminated by a threshold (3·RMSE).

High elevation accuracies of $\sigma_h=0.45 \cdot GSD$ were obtained when well defined points were manually measured in DMC images by an experienced operator (Spreckels et al., 2010).

Other investigations were carried out in (Höhle, 2009a) and (Höhle, 2009b) using UltraCam-D and DMC images. The results from these investigations will be used for estimating the impact of the four new digital aerial cameras.

In order to estimate the results with the new cameras the factors influencing the accuracy have to be known. In general, the estimates of the absolute elevation accuracy can be calculated after the formula (2).

$$\sigma_h = \frac{h}{b} \cdot \sigma_{px'} \cdot m_b \quad (2)$$

where σ_h= elevation accuracy, h = (mean) flying height above ground, b= basis, $\sigma_{px'}$=parallax accuracy related to the image, m_b= image scale figure.

The formula can be 'modernized' with regard to digital cameras:

$$\sigma_h = \frac{c}{b'} \cdot \frac{\sigma_{px'}}{pel'} \cdot GSD \quad (3)$$

When applying the factor K_{cam} the absolute elevation accuracy is obtained by means of formula (4):

$$\sigma_h = K_{cam} \cdot \sigma_{px'} \cdot GSD \quad (4)$$

where

$$GSD = pel' \cdot m_b = pel' \cdot \frac{h}{c} \cdot$$

The relative elevation accuracy is then derived by formula (5).

$$\frac{\sigma_h}{h} = \frac{\sigma_{px'}}{b'} \quad (5)$$

The image base (b') may be determined from the number of pixels in the direction of flight (s'_{fld}) assuming 60% forward overlap. The relative accuracy of parallaxes ($\sigma_{px'}$ /pel') can empirically be derived by means of equation (6).

$$\frac{\sigma_{px'}}{pel'} = \frac{b'}{c} \cdot \frac{\sigma_h}{GSD} \quad (6)$$

In (Höhle, 2009b) the relative accuracy has been derived for the DMC camera (cf. Table 2). The required value for the absolute accuracy (σ_h) has been found from comparison with accurate reference points.

b'/c b/h	GSD [cm]	spacing [m]	σ_h [cm]	$\frac{\sigma_h}{GSD}$	$\sigma_{px'}$ [µm]	$\frac{\sigma_{px'}}{pel'}$
0.31	20	3.0	17.0	0.85	3.2	0.26
0.31	10	1.6	15.0	1.50	5.5	0.46

Table 2. Results of automatic DTM derivation in open terrain using the DMC camera (c=120mm, pel'=12µm, s'_{fld}=7680pixel, b/h=0.31 at p=60%), standard overlap (p=60%, q=20%) and processing with Match-T

The mean of the relative parallax accuracy is 0.36 which corresponds to an absolute parallax accuracy of $\sigma_{px'}$=4.3µm. This value is independent from the used overlap. It should be a constant value for the DMC camera provided that the orientation of the images is error-free and the terrain is not covered with vegetation or buildings.

Other cameras may have other values. For example, in (Höhle 2009a) the parallax accuracy has been determined for the UC-D camera (c=101mm, pel'=9µm, b/h=0.27 at p=60%, s'_{fld}=7500pixel) under similar conditions (GSD=6cm, p=60%, q=20%, processing with Match-T) and the parallax accuracy has been $\sigma_{px'}$=6.0 µm or 0.67. The average from all three tests is $(\sigma_{px'}/pel')_{average}$=0.46 ≈ 0.5.

The results obtained in the above mentioned tests cannot directly be compared with results of tests using standard overlap and single models for the automated generation of elevations of natural terrain points. In the following accuracy estimation for the four new cameras it is assumed that the relative parallax accuracy is $\sigma_{px'}/pel'$=0.5 as well. The figure 2 depicts the relative elevation accuracy and Figure 3 the absolute accuracy of the elevations which the four new cameras may achieve.

From Figure 2 it can be read that the relative accuracy (σ_h / h) will improve at all of the new cameras. However, the increase in the number of pixels in the direction of flight does not mean that also the absolute accuracy improves (cf. Figure 3). Both the base/height ratio and the image scale have influence on the absolute elevation accuracy. The absolute elevation accuracy with the DMCII-250 will not be better than the results with the DMC. In order to obtain a vertical accuracy of σ_h=10cm, the DMCII-250 images have to be taken from an altitude h=1180m (corresponding to GSD=5.9cm).

When mass points for the purpose of terrain modelling are automatically determined the terrain type will have some influence. Terrain areas with buildings and vegetation have to be filtered and arising gaps have to be closed by interpolation. Errors will occur in these processes and the obtainable accuracy is then less than in open terrain. In (Höhle 2010) the DTM accuracy in built-up areas has been assessed with σ_h=16.5cm or 0.165 ‰·h. The applied DMC imagery had standard overlaps (p=60% q=20%) and a GSD of 10cm.

2.2 Coverage in object space

Due to a newly designed lens the format of the CCD in the DMCII-250 is completely used for imaging. The ground coverage at GSD=10cm is 1722m x 1466m or 2.52 km^2. The UC Eagle and the IGI-235 have about the same footprint (cf. Table 3).

A large 'footprint' is an economic factor. The flying altitude for the DMCII-250 has to be higher, which may reduce contrast and colour balancing in the images.

2.3 Increase of overlap and combined use of imagery

The accuracy can be improved when the elevations are determined from several images. This multi-ray approach is achieved by higher overlaps. The increase of the forward overlap to 80% is achieved almost without additional costs. The small frame rate of the new cameras, e.g. 1.7 seconds at the DMCII-250, makes a ground resolution of GSD=5cm possible at ground speeds of the aircraft up to 310 km/h. Figure 4 depicts an image together with the centres of four overlapping images and Table 4 shows base/height ratios for various combinations of image pairs. The elevations are derived from image pairs and have to be merged or the intersection of points is carried out simultaneously from several images. Special software for the processing is then necessary and also available from different vendors, e.g. (Erdas, 2010), (Inpho, 2011). Using several photogrammetric models and merging the derived DEMs will reduce the number of blunders and improve the accuracy. The reduction of the standard deviation may be estimated by:

$$\sigma_{h\_average} = \frac{1}{\sqrt{n_{\mathrm{mod}}}} \cdot \sigma_h$$

where n=number of models.

The accuracy is also improved by means of higher side-overlap, e.g. of q=60%. This approach increases the cost of flying. Generally, a large base line leads to differences in the perspective distortion, which may create blunders at large scale imagery. In contrast, a small base line will improve the matching accuracy, but the elevation accuracy will suffer due to the unfavourable base/height ratio. The error propagation in multi-ray photogrammetry for DEM generation needs some further investigation and practical experience. The current tests of a benchmarking may give some answers in this respect (EuroSDR, 2011).

Of advantage is also the approach with two flying heights. It may be practiced in projects where imagery is taken for built-up areas with GSD=10cm and for the whole territory with GSD=20cm and applied for orthoimage generation and stereo compilation. When using such imagery for DEM generation a 17% improvement in the standard deviation in the open land and 42% in the built-up areas was obtained when merging several DEMs (Höhle, 2009b).

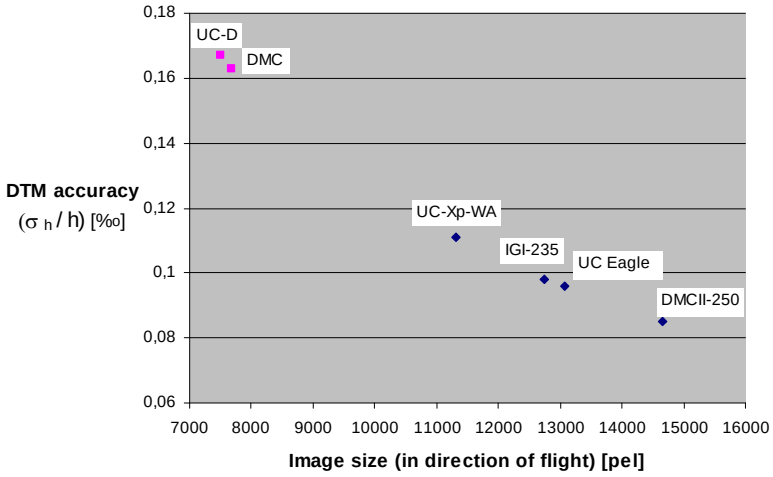


Figure 2. Relative elevation accuracy of digital aerial cameras

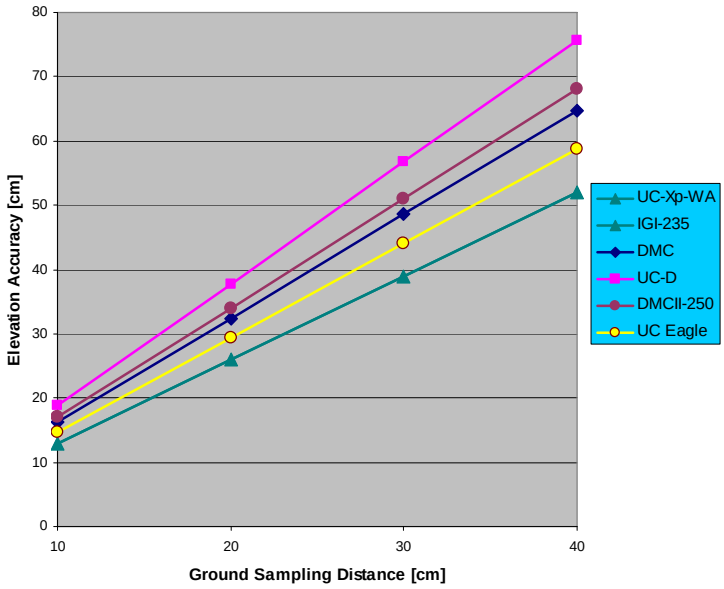


Figure 3. Absolute elevation accuracy for digital cameras as function of GSD

Features		IGI 235	UC-Xp WA	DMCII 250	UC Eagle
flying height at GSD=10 cm	[m]	1333	1167	2000	1538
image size (across flight direction)	[pel]	18500	17310	17216	20010
	[mm]	111.0	103.9	96.4	104.1
at GSD=10 cm	[m]	1850	1731	1722	2001
Image size (in flight direction)	[pel]	12 750	11 310	14 656	13 080
	[mm]	76.50	67.86	82.07	68.02
at GSD=10 cm	[m]	1275	1131	1466	1308
ground coverage at GSD=10cm	[km^2]	2.36	1.96	2.52	2.62
relative coverage	[%]	90	75	96	100

Table 3. Ground coverage of the four new digital cameras

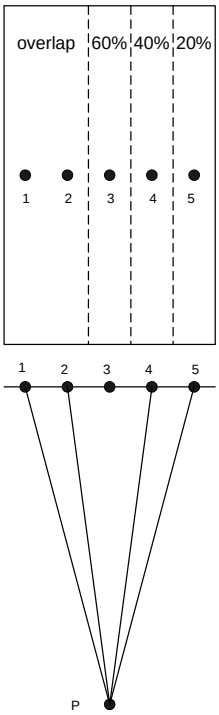


Figure 4. Multiple image pairs may be used to determine the elevation of an object point (P). For image pairs with less overlap the base/height ratio is increased and the accuracy of the derived elevations is improved.

Image pair	overlap	base/height ratio
1,5	20 %	1/1.9
1,4	40 %	1/2.5
2,5	40 %	1/2.5
2,4	60 %	1/3.7
2,3/3,4	80 %	1/7.5

Table 4. Image pair combinations and the base/height ratios of the digital large format camera UltraCam-D.

2.4 Image quality

Important for good results in geometry is also the quality of the produced image. Regarding the panchromatic camera of the DMCII-250, the resolving power of the new Carl Zeiss lens has to match the size of the small pixel. A filter removes light beyond the wavelength of λ=710nm. Colour and colour infra-red images are produced by means of pan-sharpening from four multi-spectral cameras of lower resolution and smaller camera constant. The PAN/colour ratio is 1:3.2. The forward image motion due to speed of the aircraft is compensated in all four cameras by time-delayed integration (TDI). The DMCII-250 camera has a radiometric resolution of 14bit (or 16384 intensity values) for each pixel. All this is an important condition for high parallax accuracy. Details of the radiometric post-processing are dealt with e.g. in (Honkavaara et al., 2009). The other three new cameras have similar features regarding the radiometric quality. The IGI-235 (also called Quattro DigiCAM) has interchangeable lenses with focal lengths of f=80, 100, and 150mm. The UC Eagle can also be used with a 210mm lens. DEM generation is best done with lenses of small focal length. It is important to understand that a small pixel will not necessarily result in high resolution of the image. The resolving power of the lens has to be equally high. A CCD with pixels of 5.6µm·5.6µm requires a lens with a resolution of R=1000/(2·5.6µm)=89 line pairs per mm (lp/mm). The resolution of the lens decreases from the middle of the lens to the corners. Furthermore, diffraction will occur when high f-numbers have to be used for correct exposures. The limitation of resolution of a lens due to diffraction is computed after

$$x=1.2\cdot\lambda\cdot b \qquad (7)$$

where x=smallest distance, which can be resolved, λ=wavelength of the used light, and b=f-number.

(Schwidefsky, 1976, p.48).

Assuming λ=0.575µm and b=16 then x=11µm.

The resolution of the lens at this f-number (b=16) is limited to 91 lp/mm due to diffraction. This means that exposures with f-number >16 cannot take advantage of the resolution given by a pixel size of $(5.5\mu m)^2$.

3. DISCUSSION

The recent announcement of new digital aerial cameras has started this investigation to estimate their impact on automated DEM acquisition. All four cameras have a much larger number of pixels per final output image than the cameras UC-D and DMC. It is remarkable that the 250 Megapixels of the panchromatic camera of the DMCII are achieved with one CCD array and one lens. This fact will improve the geometric accuracy because errors in the fusion of several CCDs are avoided and only one perspective centre exists. The pixel sizes in all cameras are now very small, which results in higher flying altitudes. Then the relative accuracy is better, but important for the user is the absolute accuracy for a given GSD.

The empirical tests with the first generation cameras (DMC and UC-D) revealed parallax accuracies of about 0.5 pixels in average. Applying this value for all of the new cameras the absolute accuracy will then be higher for the large-format cameras (IGI-235 and UC-Xp WA) due to their bigger base/height ratio (0.38 and 0.39 respectively). The one-chip camera (DMCII-250) comes close to the results of the large-format camera DMC. It is also advantageous that the elevations can be determined by several images due to the short frame rate in the new cameras. This multi-ray photogrammetry requires higher side overlap and special software packages. Practical experience is still necessary in order to use this approach with its full potential.

An alternative solution is the determination of DEMs from images taken at different altitudes and using the standard forward overlap (p=60%) and merging the different DEMs. The completeness, reliability and the accuracy of the DEMs can be improved by means of this approach.

The economy of DEM generation will be improved with the four new cameras. The swath width is increased either due to a larger field of view (UC-Xp WA, IGI, 235) or due to a higher flying height (DMCII-250, UC Eagle), which the small pixel or the longer focal length require. The ground resolution, however, will still be high. In order to recognize and measure small objects like manhole covers and drain gratings a high ground resolution is necessary. In the mapping industry is a tendency to use imagery with GSD smaller than 10cm.

The high resolution and image quality of the new digital cameras enable very dense point clouds at the automated DEM generation. The terrain can be modelled more accurately. Break lines, buildings and other objects can automatically be extracted more accurately and reliably.

4. CONCLUSION

Determination of elevations and DEMs will benefit from using one of the new generations of cameras. The improvements are due to a more favourable base/height ratio at the UC-Xp WA and IGI 235 cameras and at the DMCII-250 where the fusion of several CCDs is avoided. In this case the format of the image becomes nearly a square, which then matches the FOV of the single lens. The UC Eagle with its 260 Mega pixel per final output image sets a new scale regarding the amount of pixels.

Applying a relative parallax accuracy of $\sigma_{px'/pel'}$=0.5 for all four new cameras the achievable accuracy can be estimated and compared. In order to obtain a vertical accuracy of σ_h=10cm the images have to be taken with a ground sampling distance of GSD=5.9cm (DMCII-250), 6.8cm (UC Eagle), 7.6cm (IGI-235, c=80mm), or 7.8cm (UC-Xp WA).

The ground coverage of the UC Eagle is largest (100%), 96% of the DMCII-250, 90% of the IGI-235 camera, and 75% of the UC-Xp WA camera (80 mm lens) when the same ground resolution is maintained. Due to a short frame rate the multi-ray and multi-model techniques can be applied at all new cameras which will further improve the vertical accuracy. Other applications, e.g. the extraction of break lines or buildings, will benefit from the new cameras too.

Practical tests with the new cameras have to be carried out in order to confirm these theoretical considerations.

With the appearance of the four new aerial cameras the development of digital aerial cameras is not finished. The CCDs may become larger and squared and used in single chip or multi-chip designs. The image quality and other features of the cameras will improve in future as well. All this progress in the cameras and the processing tools will definitely improve the DTM/DSM generation by means of photogrammetry.

REFERENCES

EuroSDR, 2011. Benchmarking of Image Matching Approaches for DSM Computation, http://eurosdrbenchmarkofimagematching.ign.fr/ (accessed 15.5.2011)

Grenzdörffer, G., 2010. Medium Format Cameras, EuroSDR Official Publication no. 58, pp. 233-262.

Haala, N., Hastedt, H., Wolf, K., Ressl, C., Baltrusch, S., 2010. Digital Photogrammetric Camera Evaluation - Generation of Digital Elevation Models, *PFG* 2/2010, pp. 99-115.

Höhle, J., 2009a. DEM Generation Using a Digital Large Format Frame Camera, *Photogrammetric Engineering & Remote Sensing (PE&RS)*, Vol. 75, no.1, pp. 87-93.

Höhle 2009b, Updating of the Danish Elevation Model by means of photogrammetric methods, National Survey and Cadastre—Denmark, technical report series number 03, ISBN87-92107-25-7, 64 p., http://www.kms.dk/NR/rdonlyres/1C10C559-6CC9-4520-85C5-DE8659CB38A9/0/kmsrep_3.pdf (accessed 23.4.2011).

Höhle, J., 2010. Generation and Application of Digital Elevation Models, dissertation at Aalborg University, 177 p., http://people.plan.aau.dk/~jh/Articles/Hoehle_D_final.pdf (accessed 23.4.2011).

Honkavaara, E., Arbiol, R., Markelin, L., Martinez, L., Cramer, M., Bovet, S., Chandelier, L., Ilves, R., Klonus, S., Marshal, P., Schläpfer, D., Tabor, M., Thom, C., Veje, N., 2009. Digital Airborne Photogrammetry—A New Tool for Quantitative Remote Sensing?—A State-of-the-Art Review On Radiometric Aspects of Digital Photogrammetric Images, Remote Sensing, vol. 1, pp. 577-605.

Sandau, R. (editor), 2005. Digitale Luftbildkamera, Einführung und Grundlagen, Wichmann Verlag, ISBN 3-87907-391-0, 342p.

Spreckels, V. Syrek, L., Schlienkamp, A., 2010. DGPF Project: Evaluation of Digital Photogrammetric Camera Systems – Stereoplotting, *PFG* 02/2010, 14p.

Schwidefsky, K., Ackermann, F., 1976. Photogrammetrie, 7th edition, B.G. Teubner, Stuttgart. ISBN 3-519-13401-2

COMPANY LITERATURE

Erdas, 2011. LPS 2011, brochure, 16 p., http://www.erdas.com/Libraries/Tech_Docs/LPS_2011_Product_Description.sflb.ashx (accessed 17.5.2011)

IGI, 2010. Quattro DigiCAM-Modular Large Format Aerial Camera, brochure, 2p., http://www.igi.eu/brochures.html, (accessed 26.4.2011).

Inpho, 2011. Match-T DSM, brochure, 2p., http://www.inpho.de/index.php?seite=index_match-t&navigation=185&root=165&kanal=html (accessed 13.6.2011)

Intergraph, 2010. DMC® II250 CAMERA SYSTEM, brochure, 2p., http://www.intergraph.com/assets/plugins/sgicollaterals/downloads/DMCII250-CameraSystem_ProductSheet.pdf (accessed 26.4.2011).

Microsoft, 2011, UltraCamXp WA, Technical Specifications, 1p., http://www.microsoft.com/ultracam/en-us/UltraCamXpWATechnical.aspx (accessed 13.5.2011)

ACKNOWLEDGEMENTS

The author thanks the unknown reviewers for their constructive criticism.

ASSESSMENT OF RADARSAT-2 HR STEREO DATA OVER CANADIAN NORTHERN AND ARCTIC STUDY SITES

Th. Toutin [a, *], K. Omari [a], E. Blondel [b], D. Clavet [c], C.V. Schmitt [a]

[a] Canada Centre for Remote Sensing, 588 Booth Street, Ottawa, K1A 0Y7, Canada - (toutin, omari)@.nrcan.gc.ca
[b] Gismatix Inc., 1475 Cumberland Ridge Drive, Cumberland, K4C 1E1, Canada - enriqueblondel@rogers.com
[c] Centre for Topographic Information, 2144 King Street West, Sherbrooke, J1J 2E8, Canada - Clavet@.nrcan.gc.ca

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: Radarsat-2, SAR, High-resolution, Geometric modeling, DSM, Canadian Arctic

ABSTRACT:

Digital surface models (DSMs) extracted from high-resolution Radarsat-2 (R2) stereo images using a new hybrid radargrammetric modeling developed at the Canada Centre for Remote Sensing are evaluated over two Canadian northern and arctic study sites. Because the new hybrid model uses the full metadata of R2, it does not require any ground control point. The first study site in the north of Quebec is used for the scientific validation where accurate checked data (dGPS, lidar) is available. The second study site in the Arctic (steep relief and glaciated surfaces) is challenging for the operational evaluation of topographic mapping capabilities of R2. For the first study site, the bias and elevation linear errors with 68 percent confidence level (LE68) of R2 stereo-extracted DSM compared to lidar data were computed over bare surfaces: LE90 of 3.9 m and no bias were achieved. For the second study site the comparison was performed between the R2 DEM and ICESat data. A negative 18-m bias was computed and certainly result suggests a bias in the stereo-model of R2 and thus in the metadata used in the model computation because there is few temporal variation in the data acquisition (R2 and ICESat). LE68 of 28 m was obtained. However, the differential melting and thinning depending of the glaciers elevations and planimetric surging of glacier tongues with less accumulation of debris and moraines, a lower LE68 of around 20 m could be expected. In addition to evaluate the potential of R2 over ice bodies, which generally have low slope relief and because the errors are strongly correlated with slopes, other statistical results of elevation differences were also computed: LE68 of 15 m was obtained over ice fields with 0-5° slopes while a little more than 20-m over less than 30° slopes was achieved.

1. INTRODUCTION

1.1 Canadian Context

Canadian Arctic suffers from poorly-known relief. In addition, the surface state of glaciated regions is rapidly evolving due to snowfall, snow transport by wind and/or surface melt, remote sensing data (Synthetic Aperture Radar (SAR) or optical) used to retrieve the topography must be acquired with the shortest possible time interval to maximize their coherence or correlation. This is also important because the flow of the glacier (up to a few meters per day) during this time interval can bias the topographic measurement.

Stereo radargrammetry using synthetic aperture radar (SAR) data can thus be an appropriate solution, even with multi-date across-track capability, due to different SAR advantages specific to ice regions. First, the backscatter of SAR sensors is more dependent of the rugosity or the dielectric component, which enable more radiometric contrast over ice fields with supraglacial debris, rock glaciers, moraines, etc. Second, the SAR sensors are operated in all-weather conditions and not dependent of the solar illumination conditions, which thus cancelled the large shadowed areas with optical data in high latitudes. Third, the convergence of heliosynchronous orbits to North/South poles combined with a large range of look angles (20°-60°) gives thus a strong advantage to drastically reduce the temporal variations to few days between the multi-date stereo-images acquired in the highest latitudes. Fourth, the new satellite SAR sensors have now high-resolution (HR) capabilities (sub- to few meters), and are dedicated toward operational applications.

1. Due to the remote and harsh environments of the Canadian Arctic, the 3D geometric processing of SAR images should require no ground control points (GCPs) collected by users for the operational applications. A new hybrid radargrammetric model recently developed for Radarsat-2 (R2) at the Canada Centre for Remote Sensing was thus used for the stereo-modeling and the generation of digital surface models (DSM) (Toutin and Omari, 2011). In order to evaluate the performance of the process in a well-controlled environment as well in an operational environment, two Canadian northern study sites were used: the first one for the scientific validation and the second one for the operational evaluation in Arctic.

1.2 Canadian Northern Study Sites

The first study site is located north of Québec City, Québec, Canada and spans different environments: urban and residential, semi-rural and forested (Figure 2). The elevation ranges almost from 10 m in the city in the southeast to around 1000 m in the Canadian Shield in the north. The northern part is a hilly to mountainous topography (5°-30° slopes) mainly covered with forests (deciduous, conifer and mixed) while the south part is a semi-flat topography (0°-5° slopes) with urban and residential areas.

* Corresponding author.

The second study site is located in the Baffin Island, Nunavut at approximately 70° 50' N and 71° 30' W (Figure 1). There is no vegetation cover, except small plants. More than 80% is covered by ice fields with cirque glaciers (permanent ice-covered mountains), outlets and valley glaciers and glaciers tongues surrounded by spectacular fiords with 70°-90° cliffs of 500-800 m height. Bare surface mountains also with steep slopes surround the ice fields and glaciers. The valley glacier in the south-east is about 1-km wide with up-to-4° slopes surrounded by 600-m height bare surface and cirque glaciers. The elevation ranges from sea level to 1840 m and the slopes vary from 0° to 90° at fiord cliffs, illustrating a very challenging environment (in terms of land cover and relief).

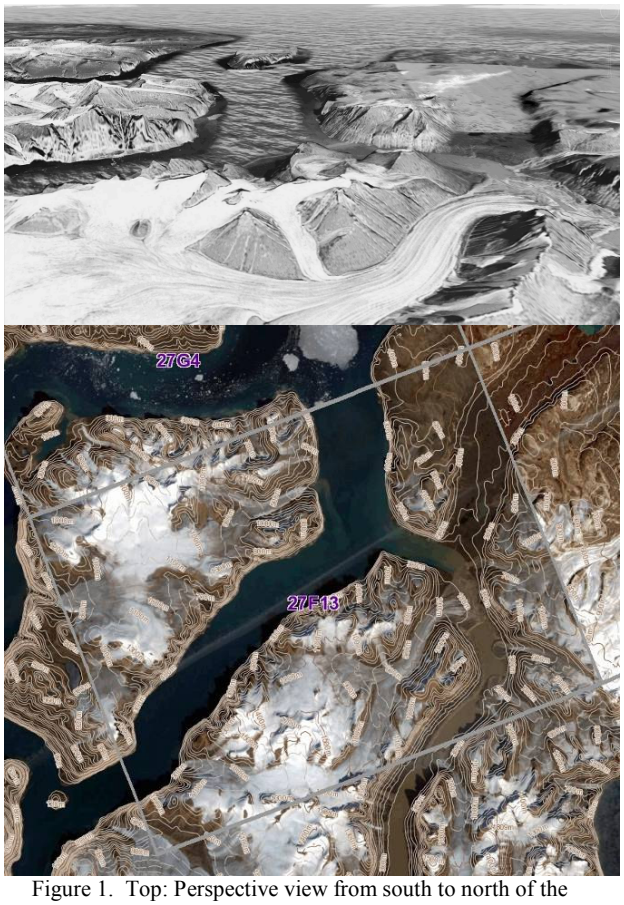


Figure 1. Top: Perspective view from south to north of the second study site, generated with Google Earth using images from TerraMetrics and WorldView. © 2010 Google and Images © 2010 TerraMetrics and DigitalGlobe.
Bottom: Ortho-rectified Landsat-7 with 1:50,000 map sheet grid

1.3 Data Set

For the first study site, the R2 SAR data set included two stereo images (20 by 20 km) acquired September 10 and 14, 2008 with the C-band ultra-fine (U) mode (1 by 1 look; 1.6-2.4 by 3 m resolution) in VV polarization from descending orbits with view angles of 30.8°-32° (U2 Figure 1) and 47.5°-48.3° (U25) at the near-far edges, respectively.

The reference cartographic data included ground points, mainly road intersections and electrical poles, collected from a differential Global Positioning System (dGPS) survey in November 2008 with 3-D ground accuracy of 10-20 cm. The collected points were used either as independent check points (ICPs) to quantify/validate the new hybrid model accuracy. In addition, 10-cm accurate cloud-point data (first echoed return) were obtained froma lidar survey collected by GPR Consultants.

For the second study site, the R2 stereo images were acquired in 2009 from descending orbit with the ultra-fine mode (U mode with 3-m resolution) in HH polarization: U12 and U26 (20 by 20 km; 1.6 by 2.5-3.0 m pixel spacing) on September 28 with 38.83°-39.84° right look angles and October 09 with 48.12°-48.93° look angles, respectively (Figure 3). The radiometry of these stereo-data are however, dominated by the geometric issues due to high relief with no vegetation cover: more severe layover in U12 over the east-oriented slopes and more occluded areas over the west-oriented slopes in U26. In fact, the south-north curved land-water boundary of the left island represents the cliff layover over the ocean, and not the "smoother" coast line of a glacial-eroded fiord, and part of the low lands along this coastline cliff thus "disappeared". On the other hand, it is almost impossible to discriminate the true water-land boundary for the opposite coastline cliffs due to the SAR shadow/occluded areas.

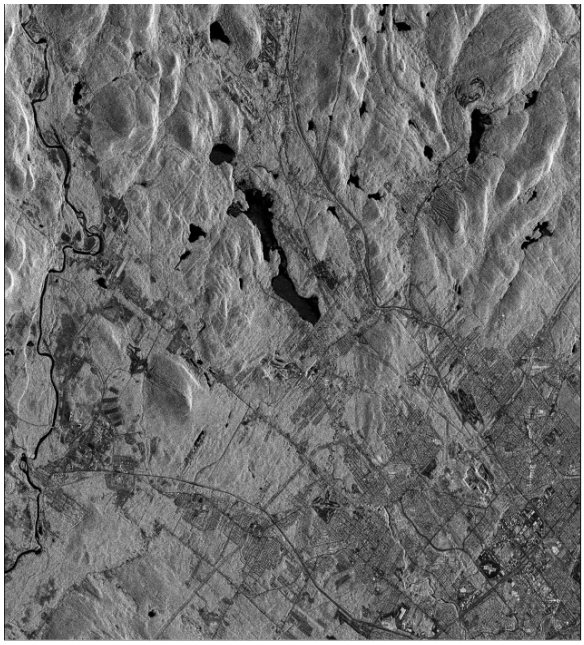


Figure 2. R2 ultra-fine mode image U2 of the 1st study site. "Radarsat-2 Data © MacDonald, Dettwiler and Associates Ltd. (2008) – All Rights Reserved" and Courtesy of CSA

We can notice on Figure 3 that all coastline cliffs and most cirque glaciers as having strong slopes (60°-90°) while the ice fields with their outlet glaciers have in general lower slopes (0-20°). However, steep 20°-90° slopes also occurred in some ice fields.

The lidar ICESat data (ascending and descending tracks) over the 27F13 map sheet was extracted from GLA14 product (L2 Global Land Surface Altimetry Data) Release 28, 29 and 31 over the full life cycle of the mission 2003-2008. Conversion to Canadian reference systems was applied to translate ICESat GLAS data into orthometric heights. ICESat data points were spatially filtered: (1) horizontally to remove redundant values

within 100-m radius, and (2) vertically within 1-arcsec grid spacing (around 25 m at 70° latitude), to remove potential elevation anomalies resulting from clouds or valley fog above 50 m above sea level.

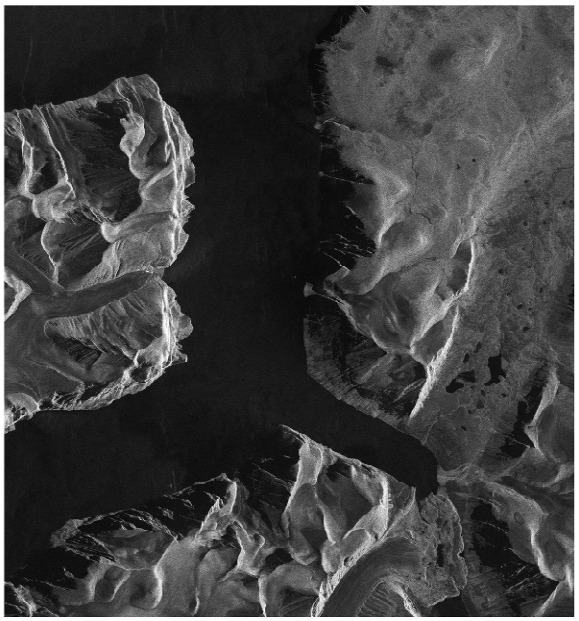


Figure 3. R2 ultra-fine mode image U12 of the 2nd study site "Radarsat-2 Data © MacDonald, Dettwiler and Associates Ltd. (2009) – All Rights Reserved" and Courtesy of CSA

2. DESCRIPTION OF THE PROCESSING STEPS

The processing steps for DSM generation with HR SAR stereo-images were previously addressed and documented (Toutin, 2010). The new hybrid Toutin's model developed for the radargrammetric processing of R-2 at CCRS does not require any GCP collected by the user (Toutin and Omari, 2011): it only uses the information in the meta-data of the images for computing its parameters. The hybrid model has been proven to be 25-cm precise (Toutin and Chénier, 2009), and the accuracy (1 σ) of the results in stereoscopy was better than one pixel with one-pixel biases in the three axes. The main processing steps are:

1. Acquisition and pre-processing of the SLC SAR images and metadata;
2. Collection of 60 ICPs from the dGPS survey;
3. Computation of hybrid models and their validation with ICPs (systematic and random errors);
4. Elevation extraction used a 7-step hierarchical grey-level image matching performed on the quasi-epipolar stereo-images and geocoding of this epipolar DSM (Ostrowski and Cheng, 2000);
5. Evaluation (systematic and random errors) of the geocoded DSMs with the lidar elevation data.

3. RESULTS AND DISCUSSIONS

Results are first related to the systematic and random errors (1) of the hybrid radargrammetric models computed over ICPs and (2) of the stereo-extracted DSMs computed over the lidar elevation data.

3.1 Radargrammetric model evaluation

Because there was no control data in the second study site, Table 1 only summarizes the results of the radargrammetric modeling computation for the first study site and data set previously described: the errors (bias and standard deviation, Std in meters) computed over 60 ICPs providing independent and unbiased evaluations of the modeling accuracy. Biases of one pixel (or half SAR resolution) or slightly worse for Y-direction are obtained. Similarly, Std results in the order of one pixel are better for X-direction. It is certainly due to the better knowledge of the range direction than the azimuth direction corresponding to the satellite displacement. Both results in Z-direction are also very good versus the SAR resolution and the same-side stereo-geometry. These results are comparable (10% difference), but a little worse in the Y-direction, to the original radargrammetric model computed with user-collected GCPs (Toutin and Chénier, 2009). On the other hand, the small lost in accuracy for the hybrid model is compensated by the gain of processing the stereo-images without GCP.

Bias (m)			Std (m)		
X	Y	Z	X	Y	Z
1.8	2.6	-2.7	0.93	1.33	2.34

Table 1. Modeling results over 60 ICPs for the 1st study site: Bias and standard deviation (Std) in metres

3.2 DSM evaluation

3.2.1 First study site: Visually, the DSM (Figure 4) with only few percent of mismatched areas is smooth and well describes the macro-topography and the macro linear trends with mountains and valleys, enhancing the structural geological framework in the northwest-southeast direction. The mountains and valleys are generally smooth, being a good representation of a Precambrian geomorphology (smoothed-glacial and eroded topography).

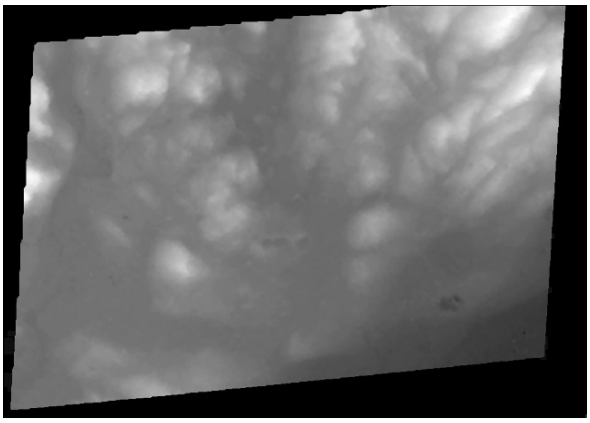


Figure 3. R2 DSM of the first study site

The quantitative evaluation was performed over the coverage of the lidar data, being on the half eastern part of DSM (Figure 3). The computed difference between R2 DSM and lidar data (Table 2) would be still representative of the overall DSM: all relief (flat to mountainous; all slopes and azimuths) and all land covers (urban, semirural, bare soils, and forested areas) of the study site were embraced.

Surfaces	Bias (m)	LE68 (m)
All surfaces	6.6	7.5
Bare surfaces	0.1	3.9

Table 2. Differences between lidar data and R2 DSM over different surface types: Bias and LE68 in metres

However, the results computed over the full lidar coverage (Table 2, 1st line) do not reflect the true DSM accuracy since the dominant source in the error budget comes from: (1) the footprint and penetration in the vegetated cover are different for both sensors (SAR and lidar); and (2) the compared stereo-SAR and lidar points are not at the same elevation in the vegetated cover (70% of lidar coverage). These errors are thus reflected in the 6.6-m bias and 7.5-m LE68. To have the true elevation accuracy, the error evaluation was performed only on bare surfaces (Table 2, 2nd line) where the stereo SAR and lidar points were at the same ground elevation. Almost no bias and 3.9-m LE68 is thus obtained. The bare surfaces were also representative of the full terrain relief because they occur not only on low lands and slopes but also in the high lands and slopes (mainly, in the northeast).

3.2.2 Second study site: The R2 DEM is displayed in Figure 5 with the ice field and glaciers boundaries (in red) and supraglacial debris and morianes boundaries (in blue). The DEM looks relatively smooth over the ice fields, even with the backscatter homogeneity in ice covered areas, mainly due to the choice of the matching parameterization. In addition, the planned R-2 acquisition at the end of the melt season increased the roughness and thus the radiometric contrast. During this period, the ice covered by supraglacial debris and dust offered its maximum degree of surface texture. It is the main reason of the few percent mismatched areas. Conversely, the DEM looks very strange along all coast cliffs displaying over 50° slopes and large geometric and radiometric differences between the two images. The combination of these geometric and radiometric distortions, which only occurred in such a challenging Arctic study site, would impede any image matching.

The quantitative evaluation was performed with ICESat lidar. The height measurements of land surface will be the prime interest for DEM quality assessments (Zwally *et al*., 2002). While the total number of ICESat data from ascending and descending orbits is limited to thousand points (Figure 5, blue and red footprints), it will be more limited because ICESat accuracy strongly degrades with slopes. Because the primary goal of ICESat is to measure inter-annual and long-term variations in the polar ice-sheet elevation and volume of Greenland and Antarctica, there were no absolute validation results over more than 5° slopes. Consequently as a function of the expected accuracy for R2 DSM, we only considered for R2 DSM evaluation the ICESat data on slopes less than 30° (Figure 5, blue footprints). Table 3 gives the computed difference between R2 DSM and ICESat data for the total DSM (715 points).and only on ice fields (386 points).

Surfaces	ICESat points	Bias (m)	LE68 (m)
All DSM	715	-17	31
Ice fields	386	-18	28

Table 2. Differences between ICESat blue footprints and R2 DSM over different surface types: Bias and LE68 in metres

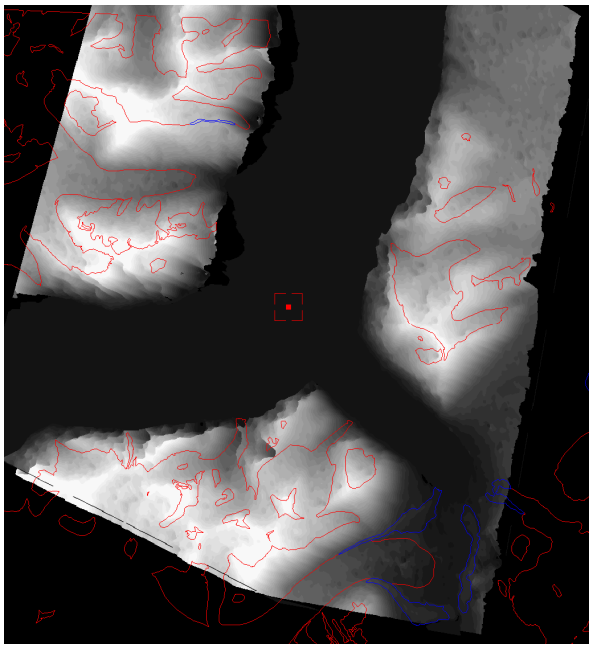


Figure 4. R2 DSM of the second study site with ice field and glacier boundaries (in red) and supraglacial debris and moraine boundaries (in blue) overlaid.

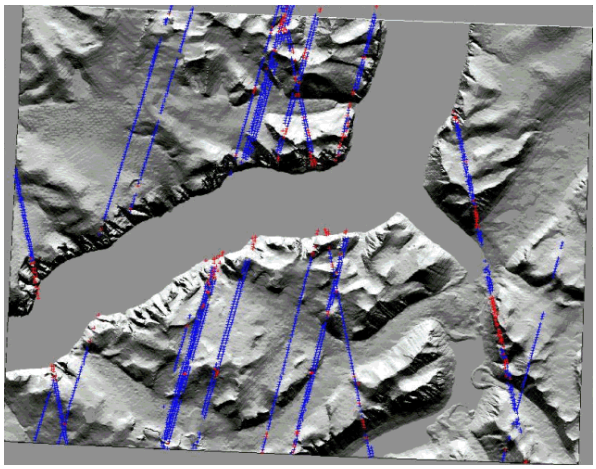


Figure 5. Ascending and descending ICESat tracks overlaid over a shaded relief image: blue and red footprints are below and over 30° slopes, respectively.

The two biases suggest a systematic error in the stereo-model of R-2 and thus in the metadata used for this stereo-model computation. The differential bias could be due to elevation change (thinning, surging) over ice fields because there is few years difference in the data acquisition. The LE68s are relatively similar but worse for all surfaces due to more steep slopes outside the ice field boundaries.

As mentioned before for the differential melting and thinning depending of the glaciers elevations and planimetric surging of glacier tongues with less accumulation of debris and moraines (Schwitter and Raymond, 1993), LE68 of R-2 DEM should thus be lower, around 20 m. In addition, most of ice bodies

have low slope relief, other statistical results of elevation differences between ICESat and R-2 DEM show that LE68 is strongly correlated with slopes: LE68 of 15 m was then computed over ice fields with 0-5° slopes while a little more than 20-m over less than 30° slopes was achieved.

4. CONCLUSIONS

A new hybrid radargrammetric model, which does not require any user-collected GCP, was evaluated with R2 stereo-data for DEM generation over two study sites in the north of Canada. The first site having accurate control data enabled DSM accuracy of 3.9 m (LE68) with no bias over bare soils to be obtained.

The second site, a challenging environment with glaciated surfaces, fjords and steep relief, was used to evaluate the mapping potential of the method in the Canadian Arctic without control data. In this remote and harsh environment, DSM LE68 of 28 m with large bias (-18 m) was achieved over ice field. This major part of this bias is certainly due to a systematic error in the metadata and partially due the ice thinning.

While other methods using optical and SAR systems could achieved similar and even better results, this application demonstrated the capability of R2 to generate DSM with better than 20 m LE68 without collecting control data over ice bodies depending of the terrain slopes (0-30°) or around 15 m over ice sheets (slopes less than 5°). This new method increases the applicability of R2 to remote and harsh environments. The slight loss in accuracy when using dGPS is then compensated by the gain of no control data.

Acknowledgements

The authors would like to thank Paul Briand and the Canadian Space Agency for supporting and financing this research under their SOAR and GRIP programs. They also thank Dr. Philip Cheng and PCI for the adaptation of CCRS algorithms in OrthoEngine^{SE} of Geomatica.

References:

Toutin, Th., 2010. Impact of RADARSAT-2 SAR Ultrafine-Mode Parameters on Stereo-Radargrammetric DEMs. *IEEE Trans. Geosc. Remote Sens.* 48(10), pp.3816-3823.

Toutin, Th. and R. Chénier, 2009. 3-D Radargrammetric Modeling of RADARSAT-2 Ultrafine Mode: Preliminary Results of the Geometric Calibration. *IEEE-GRSL*, 6(2), pp.282-286 & 6(3), pp. 611-615.

Toutin, Th. and K. Omari, 2011. A new hybrid modeling for geometric processing of Radarsat-2 data without user GCP. *Photogram. Eng. & Remote Sens.*, 77(6), pp. 601-608.

Ostrowski, J.A. and P. Cheng, P., 2000, DEM extraction from stereo SAR satellite imagery. *IEEE Proc. Geosc. Remote Sens., IGARSS 2000*, Honolulu, Hawaii, Vol. 5, pp. 2176 – 2178.

Schwitter, M.P. and Raymond, C.F., 1993. Changes in the Longitudinal Profiles of Glaciers during Advance and Retreat. *Journal of Glaciology*, Vol. 39, No. 133, pp. 582-590.

Zwally, H.J., B. Schutz, W. Abdalati, J. Abshire, C. Bentley, A. Brenner, J. Bufton, J. Dezio, D. Hancock, D. Harding, T. Herring, B. Minster, K. Quinn, S. Palm, J. Spinhirne, and Thomas, R. 2002. ICESat's laser measurements of polar ice, atmosphere, ocean, and land. *Journal of Geodynamics*, 34(3–4), pp. 405–445.

STREET REGION DETECTION FROM NORMALIZED DIGITAL SURFACE MODEL AND LASER DATA INTENSITY IMAGE

T. S. G. Mendes [a, *], A. P. Dal Poz [b]

UNESP, São Paulo State University, Brazil
[a] Cartographic Sciences Graduate Program - tatisussel@gmail.com
[b] Department of Cartography - aluir@fct.unesp.br

KEY WORDS: Street region detection, laser scanner data, normalized Digital Surface Model, intensity image, image processing techniques

ABSTRACT:

The urban road network extraction process can be simplified by firstly detecting regions corresponding to streets, allowing a substantial reduction of the search area. As a result, the extraction process is benefited in two aspects: the computational complexity and the reliability. This paper aims at detecting street regions using only data obtained by Laser Scanner Systems. A sequence of standard image processing techniques is used to process height and intensity laser scanner data. A normalized Digital Surface Model is derived from height laser scanner data, from which regions corresponding to aboveground objects (mainly trees and buildings) are detected. Next, detected tree regions are eliminated from the aboveground regions, remaining only buildings. Then, morphological operators are applied in order to obtain elongated street ribbons and homogeneous block regions. Street regions are also detected in the intensity image. The results obtained from the radiometric and geometric laser scanner data are combined, allowing the elimination of non-street regions and the improvement of the geometry of region boundaries. The experimental results showed that the methodology proved to be efficient to detect street regions.

1. INTRODUCTION

Automated urban road network extraction from digital images is an extremely complex task, since in urban environments the scenes involve a lot of objects that interact with streets. This problem can be simplified by detecting previously regions corresponding to street (RoI – Region of Interest), resulting in a considerable reducing of the search area. As a result, the computational effort and the reliability of the extraction process are improved. The use of only aerial or orbital images in the extraction process can suffer some limitations, mainly related to the similarity of the spectral response between the street and the other objects (e.g. buildings with gray roof). In order to overcome this problem, other data sources can provide additional information to support the extraction process and improve the results, as can be found in Hinz & Baumgartner (2003) that integrated high-resolution images with a DSM (Digital Surface Model) and in Zhang (2004) that integrated colour images with pre-existing spatial data base.

The use of the information from the laser scanner data can contribute as additional information. Related works can be found in Hu et al. (2004) and Zhu et al. (2004), which integrated aerial image and laser data and also in Tiwari et al. (2009) that integrated IKONOS images and altimetry laser data.

Some methodologies for road network extraction used only information obtained from laser scanner. Alharthy & Bethel (2003) used laser pulse intensity to detected candidate pixels corresponding to road and height data to removed noise. In Clode et al. (2004) laser points are classified into points of road and non-road. The classification is based on a height range where the roads can be found and a range of laser pulse intensity according to the type of material of the roads.

Streets in laser scanner data present peculiar characteristics. In normalized Digital Surface Model (NDSM), generated from 3D point cloud, the streets have homogeneous height values and are free of the occlusions caused by shadows of objects. In intensity image they appear in dark gray value, due to the low reflectance (approximately 17%) in relation of the laser pulse (Wehr & Lohr, 1999). On the one hand, the streets can be easily identified but, on the other hand, several objects (e.g. vegetation) present similar responses.

This paper presents a sequence of image processing techniques applied in both laser scanner data - intensity and NDSM images – whose goal is the detection of street regions. An important aspect of this work is the exploration of complementary characteristics of streets in these data sources.

2. METHODOLOGY

The flowchart of the proposed methodology is present in Figure 1. A 3D point cloud and intensity images, both obtained by Laser Scanner Systems, are used as input data. From the point cloud data is generated a NDSM image, which is the basis for detecting elongated street regions and homogeneous block regions. Candidate street regions are detected using the intensity image. The results are combined for the detection of the street regions.

* Corresponding author.

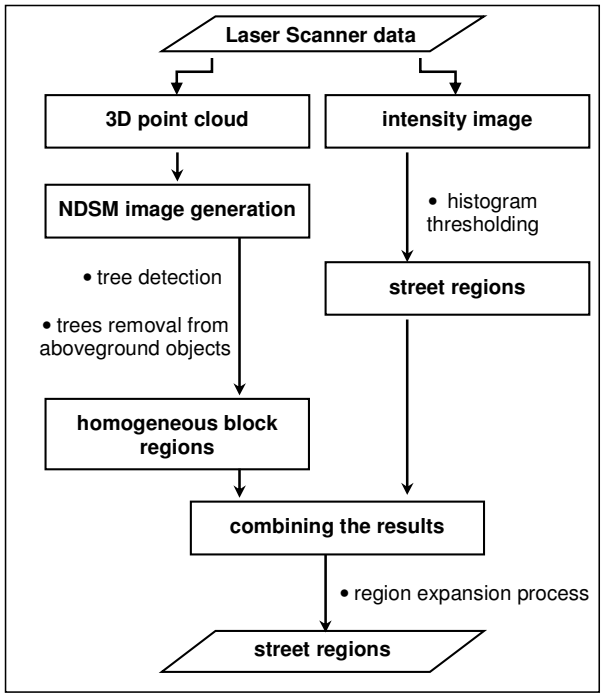


Figure 1. Proposed methodology flowchart.

2.1 Material

The study area is an urban area of Curitiba, Brazil. It is a residential area characterized by buildings, paved streets and vegetation of different sizes. The laser scanner data for this region were obtained using Optech ALTM System.

The experiments were performed using software HALCON 7.1 developed by MVTec (Machine Vision Technology).

2.2 NDSM Image Generation

Initially, the point cloud data is interpolated to generate a regular grid (DSM) using the nearest neighbour interpolator with resolution of 0.3 meters. Points representing the ground surface are collected in the DSM and another grid is generated, i. e. the DTM (Digital Terrain Model). For a representation of aboveground objects on a flat surface, a NDSM is obtained by subtracting the DTM from the corresponding DSM. The Figure 2 shows de image representing the NDSM.

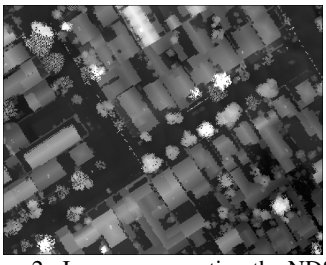


Figure 2. Image representing the NDSM.

2.3 Trees Detection from the NDSM Image

Regions corresponding to trees are detected based on texture and shape that these objects present in the NDSM image. For trees detection are necessary some steps, which are described in the following subsections.

2.3.1 Texture Filter: Trees are heterogeneous regions in NMDS image and present a texture that differentiates them from buildings (see Figure 2). Using a standard-deviation texture filter, heterogeneous regions are enhanced. This filter calculates the standard deviation of gray levels within a rectangular mask, whose dimensions (height and width) are parameters that must be provided. This operation results in a standard deviation-valued image, which is presented in Figure 3.

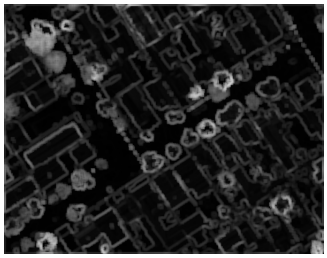


Figure 3. Standard deviation image derived from the NDSM.

2.3.2 Histogram Thresholding: In a standard deviation image (Fig. 3), the edges and trees are enhanced. Through the histogram analysis of this image is possible to choose a threshold value and to isolate such regions applying a histogram thresholding. However, not only trees but also some enhanced building contours are detected, as can be seen in Figure 4.

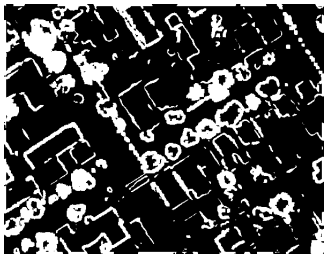


Figure 4. Regions enhanced detected from standard deviation image.

2.3.3 Morphological Operators: To eliminate non-tree regions obtained in previous step, it is assumed that trees usually have circular shape, while other objects have different shapes. Thus, morphological operators of opening and closing are applied using a disk shaped structuring element. The opening operator removes isthmus and islands, if these regions were smaller than structuring element. The closing operator fills gulfs and holes of regions smaller than structuring element. Figure 5 shows the contour of the tree regions obtained overlaid on the NDSM image.

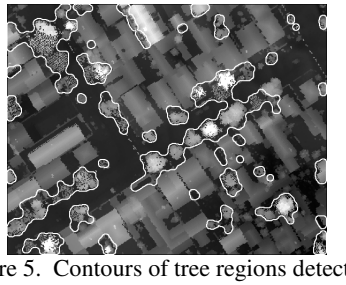


Figure 5. Contours of tree regions detected.

2.4 Trees Removal from Aboveground Objects

In a NDSM image, the street have gray level values near zero, while the aboveground objects have gray level values according to their true heights, as can be seen in Figure 2. Based on these characteristics it is possible to separate the aboveground objects through a histogram thresholding, resulting in the binary image, where the aboveground objects are represented in white, while the ground is represented in black, as showed in Figure 6.

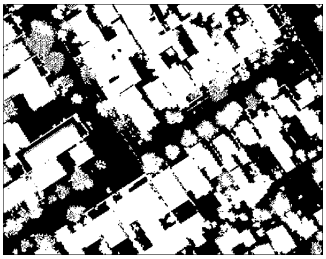


Figure 6. Binary image representing the aboveground objects.

Trees removal from aboveground objects is performed by a difference operation between the aboveground regions previously obtained (Fig. 6) and the trees detected (Section 2.3). Thus, only regions corresponding to buildings remain. Consequently, the regions that represent blocks become irregular and with holes inside (Fig. 7).

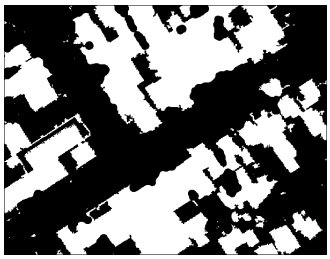


Figure 7. Trees removed from aboveground objects.

2.5 Homogeneous Block Regions

In order to obtain blocks as homogeneous regions and then obtain elongated and homogeneous street ribbons, morphological operators are applied in the previous result. The operators of dilation and erosion, both using a disk shaped structuring element, generate more regular regions, since the dilation operator fills regions smaller than the structuring element and the erosion operator clear regions smaller than the structuring element. Figure 8 shows streets as elongated ribbons, without occlusions caused by trees, and blocks as homogeneous regions.

Figure 8. Elongated street ribbons and homogeneous blocks.

2.6 Street Detection from Intensity Image

Taking advantage of the street characteristics in the intensity image, it is possible to detect candidate regions corresponding to streets applying a histogram thresholding in the intensity image. The result is a binary image representing the street regions in white and the background objects in black. Figure 9 shows the intensity image (Fig. 9(a)) and the resulting binary image (Fig. 9(b)).

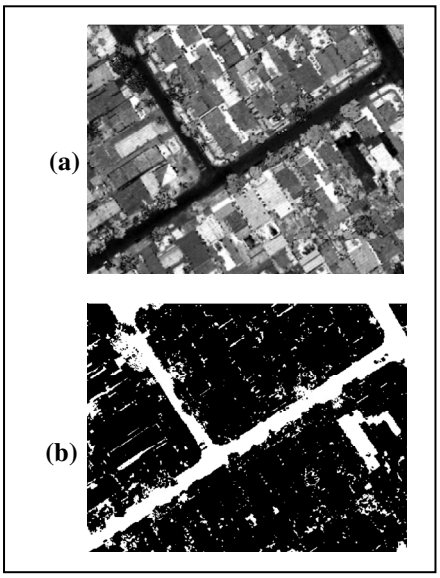


Figure 9. (a) Intensity image (b) Street regions detected by histogram thresholding.

The intensity image presents some objects that have radiometric responses similar to streets, and thus they are also detected in the histogram thresholding process.

2.7 Street Region Detection Combining the Results Obtained from Intensity and NDSM Images

The homogeneous regions representing the blocks obtained from the NDSM (Section 2.5) are combined with the candidate street regions obtained from an intensity image (Section 2.6). For this, block regions are expanded iteratively until they touch the street regions, called the forbidden areas. The expanding region operator works by adding or removing a pixel strip to a region. In this process, the non-street regions that appear inside the blocks (see Fig.9) are removed. Application of this process is shown in Figure 10, which represents the blocks, street and background regions in white, grey, and black, respectively. Figure 10(a) shows the regions before the expanding process. Expanded block regions are presented in Figure 10(b).

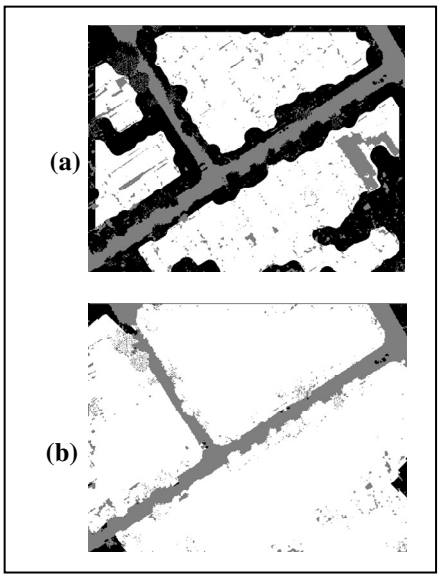


Figure10. (a) Block regions before the expanding region process. (b) Expanded block regions.

In order to obtain more regular block region boundaries, morphological operators of opening and closing are applied. We used a disk shaped structuring element. An example of result is shown in Figure 11(a). Figure 11(b) presents obtained block regions boundaries overlaid on the intensity image. Please, observe that boundary is very near to true block boundary.

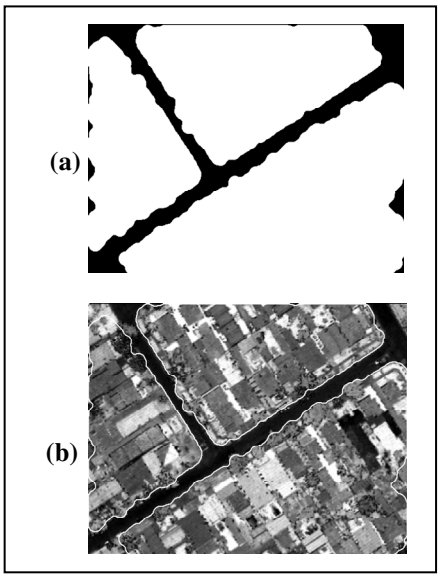


Figure 11. (a) Block regions after application of the morphological operators. (b) Block regions boundary overlaid on the intensity image.

The street region can be obtained as the complement of block regions detected previously, as shown in Figure 12. Note that street regions are successfully detected.

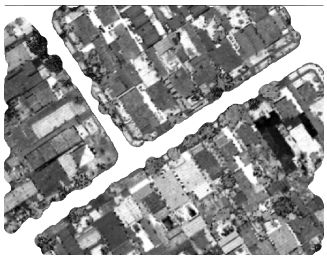


Figure 12. Result of the methodology for detect street regions overlaid on the intensity image.

3. RESULTS

In addition to results presented along with the methodology, more experiments were performed using another patch of the available laser data (see Figures 13, 14 e 15).

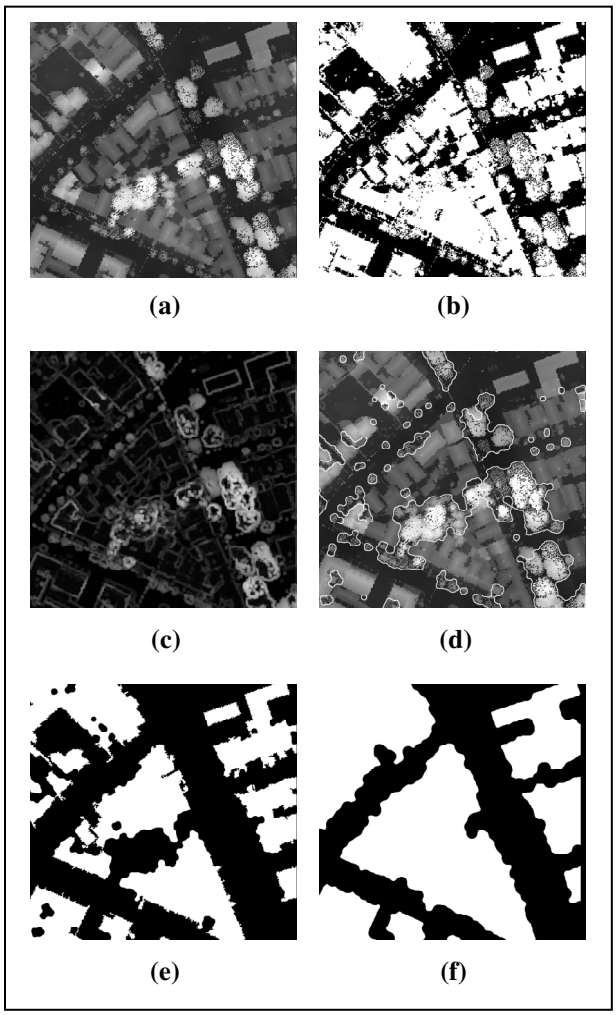


Figure 13. Results obtained from the NDSM image. (a) NDSM image. (b) Binary image representing the aboveground objects. (c) Standard deviation image derived from the NDSM. (d) Detected tree contours overlaid on the NDSM image. (e) Trees removed from aboveground objects. (f) Elongated street ribbons and homogeneous block regions.

Figure 13 presents the results obtained using the NDSM image. Figure 13(a) shows the NDSM image generated from the point

cloud data. Note that nearly the right boundary of the image there are several trees occluding almost completely the street. The aboveground objects detected by histogram thresholding are shown in Figure 13(b). Figure 13(c) shows the standard-deviation image derived from NDSM, with enhanced trees along the streets and inside the blocks. The building edges also are enhanced. Figure 13(d) presents the contours of trees detected, showing that trees are successfully detected. The difference between aboveground objects and tree regions detected is shown in Figure 13(e). The streets (black elongated ribbons) and blocks (white homogeneous regions) are presented in Figure 13(f), which also shows that the holes inside de blocks are filled.

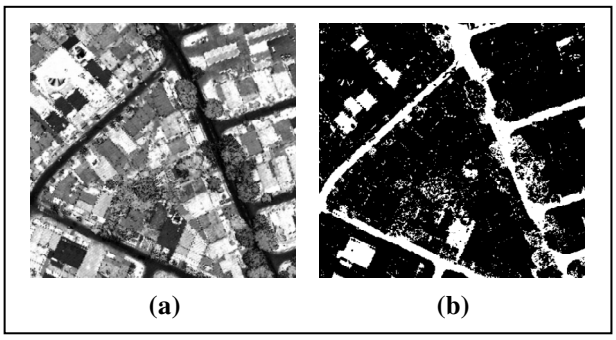


Figure 14. Results obtained from the intensity image. (a) Intensity image. (b) Binary image representing the street regions.

Figure 14(a) shows the results obtained using the intensity image. The result of the histogram thresholding is shown in Figure 14(b). Note several non-street regions, which are detected due to their similar radiometric response to our object of interest (streets).

Figure 15(a) presents the block regions (in white) obtained from the NDSM image and elongated regions (in gray) obtained through the application of the histogram thresholding to the intensity image. The elimination of the non-street regions is carried out by applying the expanding region process to these regions. Figure 15(c) shows the blocks region boundaries regularized by the application of the morphological operators. Detected street regions are overlaid on the intensity image (see Figure 15(d)). Again, the methodology proved to be efficient in detecting streets occluded by several trees.

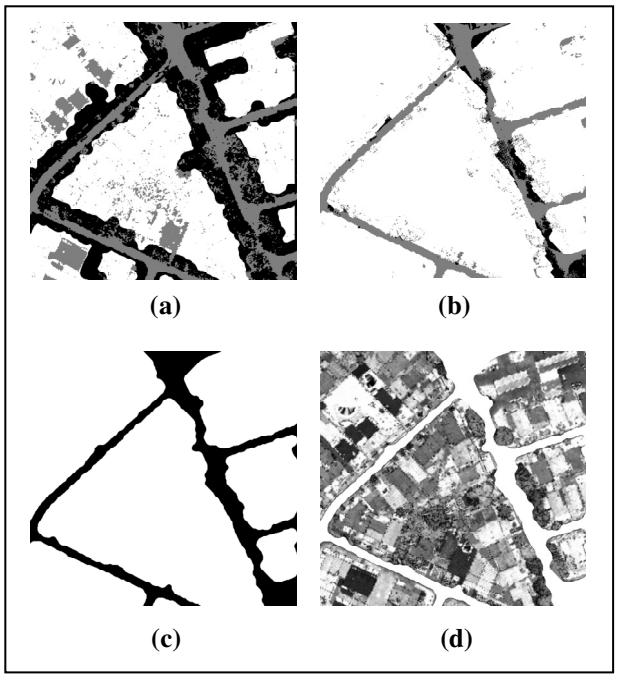


Figure 15. Results obtained combining the results generated from the radiometric and geometric laser data. (a) Block regions before the application of the region expansion process. (b) Expanded block regions. (c) Block regions after application of the morphological operators. (d) Detected street regions overlaid on the intensity image.

4. CONCLUSION

This paper presented a methodology for street region detection by using the following laser scanner data: the NDSM image generated from height data and the intensity image. The steps of the methodology were described and two experiments were presented. From a visual analysis of the obtained results it is possible to conclude that street regions were successfully detected.

Refinements and improvements are planned, mainly to regularize street region contours detected by the proposed methodology.

REFERENCES

Alharthy, A. & Bethel, J. 2003. Automated Road Extraction from LIDAR Data. In: *Proceedings of ASPRS*. Anchorage.

Clode, S., Kootsookos, P. & Rottensteiner, F. 2004. The automatic extraction of roads from LIDAR data. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science*. Istanbul, Vol. XXXV, pp. 231-237.

Hinz, S. & Baumgartner, A. 2003. Automatic extraction of urban road networks from multi-view aerial imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58(1-2), pp. 83-98.

Hu, X., Tao, C. V. & Hu, Y. 2004. Automatic road extraction from dense urban area by integrated processing of high resolution imagery and LIDAR data. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.* Istanbul, Vol. XXXV, pp. 320-325.

Tiwari, P. S., Pande, H. & Pandey, A. K. 2009. Automatic urban road extraction using Airborne Laser Scanning/altimetry and high resolution satellite data. *Journal of the Indian Society of Remote Sensing*, 37(2), pp. 223-231.

Wehr, A. & Lohr, U. 1999. Airborne LASERscanning - an introduction and overview. *ISPRS Journal of Photogrammetry and Remote Sensing*. 54 (2-3), pp. 68-82.

Zhang, C. 2004. Towards an operational system for automated updating of road databases by integration of imagery and geodata. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58, pp. 166-186.

Zhu, P., Lu, Z., Chen, X., Honda, K. & Eiumnoh, A. 2004. Extraction of city roads through shadow path reconstruction using LASER data. *Photogrammetric Engineering and Remote Sensing*. 70 (12), pp. 1433-1440.

ACKNOWLEDGEMENTS

This paper is part of a Ph. D. research supported by the Brazilian National Agency for Science and Technology (CNPq). The authors thank the LACTEC (Curitiba-Pr, Brazil) for providing the data for this research.

USING FULL WAVEFORM DATA IN URBAN AREAS

B. Molnar [a,b] *, S. Laky [c], C. Toth [a]

[a] The Center for Mapping, The Ohio State University
470 Hitchcock Hall, 2070 Neil Avenue, Columbus, OH 43210
toth@cfm.ohio-state.edu
[b] Department of Photogrammetry and Geoinformatics
Budapest University of Technology and Economics, Muegyetem rkp 3., Budapest, H-1111, Hungary
molnar.bence@fmt.bme.hu
[c] HAS-BME Research Group for Physical Geodesy and Geodynamics
Budapest University of Technology and Economics, Muegyetem rkp. 3, Budapest, H-1111, Hungary
laky.sandor@freemail.hu

Commission I, WG I/2

KEY WORDS: LiDAR waveform, classification, Self-Organizing Map (SOM), Bayes classifier, urban area

ABSTRACT:

In this paper, the use of waveform data in urban areas is studied. Full waveform is generally used in non-urban areas, where it can provide better vertical structure description of vegetation compared to discrete return systems. However, waveform could be potentially useful for classification in urban areas, where classification methods can be extended to include parameters derived from waveform analysis. Besides common properties, also sensed by multi-echo systems (intensity, number of returns), the shape of the waveform also depends on physical properties of the reflecting surface, such as material, angle of incidence, etc. The main goal of this investigation is to identify relevant parameters, derived from waveform that are related to surface material or object class. This paper uses two waveform parameterization approaches: Gaussian shape fitting and discrete wavelet transformation. The two classification methods tested are: supervised Bayes classification and unsupervised Self-Organizing Map (SOM) classification. The results of these methods were compared to each other and to manual classification. The initial conclusion is that, though waveform data contains classification information, the waveform shape by itself is not enough to perform classification in urban regions, and, consequently, it should be combined with the point cloud geometry.

1. INTRODUCTION

Most modern LiDAR system have the capability to acquire full waveform LiDAR data besides the discrete returns with intensity data. Waveform data is quite useful to distinguish tree species or provide better biomass description. The question is whether waveform can also provide useful information for object classification in urban regions. Since waveform parameters, such as the commonly used intensity, highly depend on the properties of the surface, classification can be potentially performed on them. First, the properties of the waveforms which are typical for features should be identified. In this study, two parameterizations are used: Gaussian shape fitting and discrete wavelet transformation. The methods with special extensions are described in Section 2. An important part of the research is to find typical class-specific parameters that are independent from the commonly used parameters, such as intensity or number of returns. Next, classification performance should be evaluated based on actual LiDAR data. In this investigation, two classification methods were tested: the naive Bayes supervised classification method (Green, 1995) and the unsupervised Self-Organizing Map (SOM) (Kohonen, 1990) , detailed in Section 3. The four combinations of parameter extraction and classification algorithms were tested on a LiDAR dataset acquired by an Optech ALTM 3100 sensor over an area near Dayton, Ohio, USA, shown in Figure 1. The selected area represents a typical suburban environment, including a mix of vegetated areas and man-made objects. Four object classes were defined: grass, tree, roof and pavement. All the four methods were compared to each other with respect to classification performance, described in Section 6. Finally, a validation with manually classified points was performed. All the data processing and analyses were carried out in the GNU Octave open source software environment.

Figure 1: Test area, Dayton, Ohio

2. WAVEFORM PARAMETERS

The classification methods generally require discrete well-structured input values. Since the waveforms are really different for each reflection, they cannot be directly used as input for a classification procedure. Ideally, waveforms have to be described by parameters without any loss of the material-specific information. There are many ways to model waveform (Duong *et al*., 2006)s; in this study, one typical procedure and a new method were tested. Note, that for the analysis only the return signal was used (the shape of the outgoing pulse was not considered).

2.1 Generalized Gaussian Fitting

For the purpose of determining the shape-specific waveform parameters to be used in the classification, a two-step peak detection and peak parameter extraction method were used. In the first step the number of peaks is determined by inspecting the second derivative (the curvature) of a cubic smoothing spline fitted to the waveform data. In the second step generalized Gaussian functions are fitted to the waveform data (number of the fitted functions depends on the number of peaks detected in the previous step) using the Levenberg–Marquardt algorithm (Chauve *et al*., 2007). The generalized Gaussian function used here can adjust to the translation, magnitude, pulse width, flattening and skewness of the waveform, see Figure 2. The translation represents only the location of pulse in the LiDAR data, therefore, not used for classification.

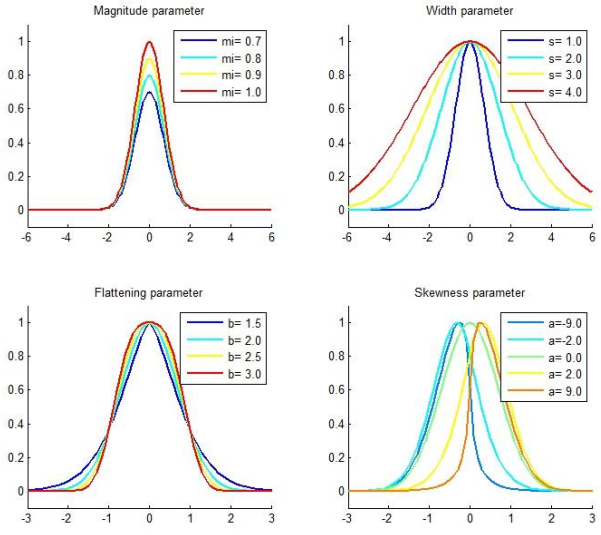


Figure 2: Four parameters of the generalized Gaussian curve

In addition to the four shape parameters, two more parameters were selected. The motivation was though they describe the waveform in a reasonable way, yet the feature specific properties do not appear really dominant outer the magnitude parameter, as seen in Figure 3. The magnitude describes the intensity value gathered by traditional scanners, that's why not so interested in this investigation. The additional parameters are expected to provide additional information for the classification, as they rely more on the waveform shape.

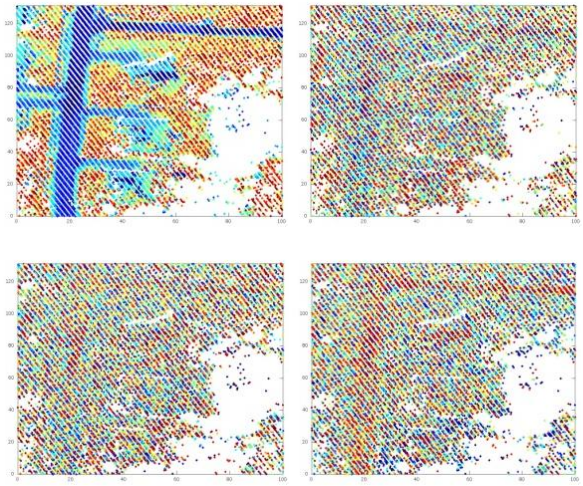


Figure 3: Magnitude, width, flattening, and skewness parameters in the test area

The 'penetration' parameter is calculated, as the number of discrete samples over a previously defined threshold. In our case the threshold was chosen to be 33. The waveforms returned from a pavement have typically lower intensity value than one from penetration. Also, this parameter better separates the vegetation (grass and tree) from the pavement and roof, as shown in Figure 4.

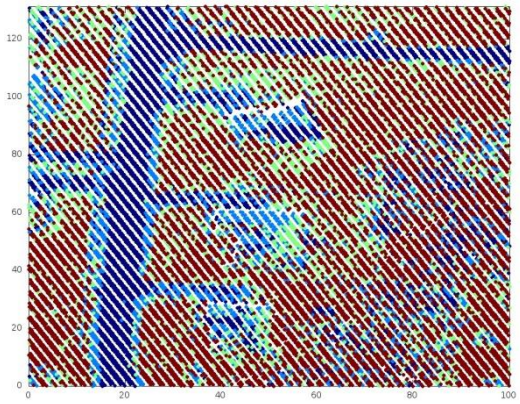


Figure 4: Penetration parameter in the test area

The classification based only on the three parameters has difficulty in separating roofs and pavement. The source of the problem is likely the different surface normals for roofs. The second additional parameter describes the residuals of the Gaussian fitting. The standard deviation of the fitting error is typical for the selected four main classes, see Figure 5.

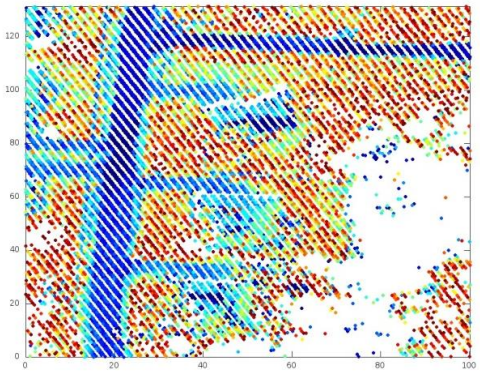


Figure 5: Fitting error in the test area

2.2 Discrete Wavelet Transformation

The waveform signal can also be transformed by Discrete Wavelet Transformation (DWT), resulting in a well compressed and structured dataset. Since the waveform has local correlation, the higher order DWT coefficients can be usually discarded. The two-pulse waveform example in Figure 6 shows that the first 18 wavelet coefficients are sufficient to preserve the waveform, and can be potentially used for classification. The CDF 3/9 wavelet transformation provides a good representation of the waveform with good compression performance (Laky *et al*., 2010). In our investigation, the WaveLab toolbox was used (Buckheit and Donoho, 1995).

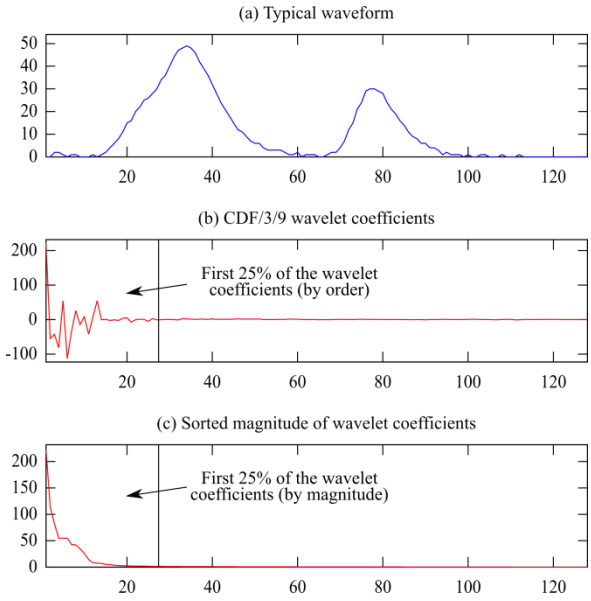


Figure 6: Wavelet coefficients

3. CLASSIFICATION METHODS

3.1 Self-Organizing Map

Automatic classification can be performed by Kohonen's Self-Organizing Map algorithm (SOM) (Kohonen, 1990). SOM is an unsupervised method and has very flexible parameterization with good performance in handling non-linear mapping problems (Zaletnyik *et al*., 2010). Our implementation used the SOM_PAK , (Kohonen *et al*., 1996).

3.2 Bayes Classifier

The second classifier tested was a naive Bayes classifier. Using a training set, the relative frequencies of the parameters falling into specified intervals for each class are calculated; i.e., the continuous parameters are discretized by binning, and then the empirical histograms of the parameters for each class are calculated. The relative frequency of the categories occurring among the training waveforms is also calculated. The class for a specific waveform is then selected to be the class that maximizes the probability

$$p(C = c) \prod p(F_i = f_i | C = c) \qquad (1)$$

where C is the class,
F_i are the classification parameters,
$p(F_i = f_i | C = c)$ is the probability of a given classification parameter to be in a given bin for a given class.

4. CLASSIFICATION TESTS

To perform comparative performance evaluation, the introduced algorithms were tested using a LiDAR dataset, acquired over Dayton, Ohio. Note that all classification methods had a post processing step with mode filtering to avoid class speckle.

4.1 Fitting and SOM

The algorithm of this method is based on the fitting parameters, especially on the pulse width, flattening, skewness, fitting error and penetration. SOM classification with a rectangle topology and 2x2 dimensions was applied on the parameter set. The parameter calculation assumes that only single peak echoes are processed; note that the multi-echoes were added during the post processing to the tree class. Furthermore, the range differences were calculated and local high points were classified as roof. This step improves the separation of roof and pavement; however, this means that not only the waveform is used for classification.
The crucial area is the sidewalk and the grass belt along the street. Fitting and SOM classifies this area as a pavement, so the pavement area is larger than in reality (Figure 7).

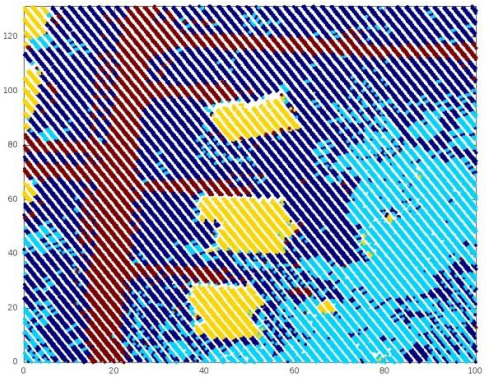


Figure 7: Fitting and SOM

4.2 Fitting and Bayes classification

The waveforms with one peak are processed by the shape fitting algorithm. In addition to the parameters introduced in the previous section, the range differences were also used to improve the separation of roof and pavement. Multi-echoes are classified as trees and local high points as a roof, similar to the method in section 4.1.
This algorithm made some differences on the sidewalk; however this area gets defined in the wrong class (Figure 8). The runtime of Bayes classification is about the same as the SOM.

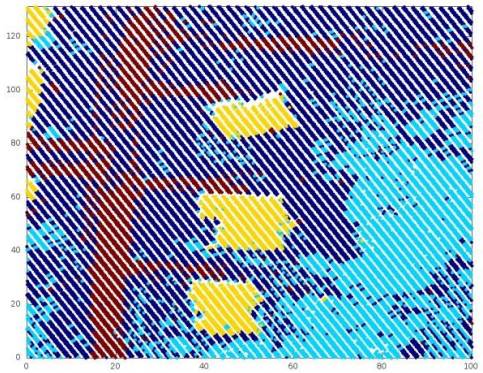


Figure 8: Fitting and Bayes

4.3 Wavelet and SOM

The benefit of wavelet is that single-echo and multi-echo waveforms can be processed in the same classification step. The first 18 wavelet coefficients were used as the input to SOM. The topology is a rectangle and 2x2 dimensions were used as well as in section 4.1. Figure 7 and Figure 9 show that SOM doesn't recognize the sidewalk and grass strip.
The separation of pavement and roofs give less reliable results. In the roof area, there are both pavement and roof points. The easiest way to improve this classification is to use range differences or to examine the height distribution in the area.

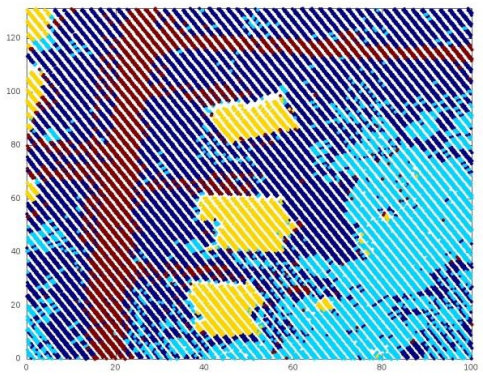


Figure 9: Wavelet and SOM

4.4 Wavelet and Bayes classification

In this case the classifier was also applied to the first 18 wavelet coefficients. The training set for Bayes classification has had 427 points about the same distribution as the final classes. The sidewalk has some usable information (Figure 10). The experiences suggest that substantial differences exist on the SOM and Bayes classifiers in this area and there are no significant differences at other regions.
This result is very similar to the "Fitting and Bayes classifier" method of Section 4.2. This shows that both sets of input parameters have the same information included and the classifier has higher impact. The other reason is the impact of non waveform based classification components.

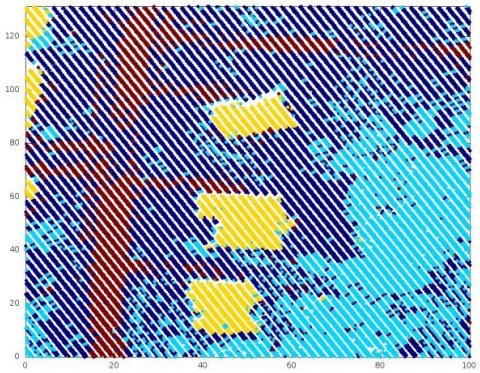


Figure 10: Wavelet and Bayes

4.5 Methods summary

Table 1 shows a summary of the used input parameters in the four methods.

	Fitting parameters	Wavelet parameters	Range differences	Multi-echoes
Fitting and SOM	5		1	1
Fitting and Bayes		18	1	1
Wavelet and SOM	5		1	1
Wavelet and Bayes		18	1	1

Table 1: Used parameters for methods

5. DISCUSSION

5.1 Improving classifications

The classification based solely on waveform parameters seems to give insufficient results. The method seems to be good for separating pavement and grass; however there are some difficulties differentiating between pavement and roof. The source of this difficulty is the different incidence angle on the roofs. This is why the range difference was used in all methods for post processing.

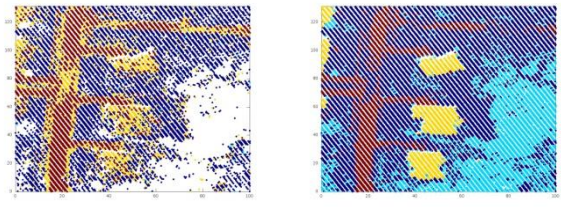


Figure 11: Change in the classification result caused by using range differences and adding multi-echo waveforms

5.2 Result Differences

To get comparative performance of the four procedures, the results were compared to each other. There were 15617 points in the selected area, and for the Bayesian classifiers, 427 training points were used. The distribution of classes was the following: 53% grass, 26% tree, 7% roof, 14% pavement, see Table 2;typical for a sparsely populated suburban area.

	Grass	Tree	Roof	Pavement
Fitting and SOM	8050	3876	1180	2511
Fitting and Bayes	8710	3955	1100	1853
Wavelet and SOM	8119	3906	1112	2411
Wavelet and Bayes	8499	4224	1143	1782
Average	8245	3990	1750	2139

Table 2: Number of points in classes

The four methods produced quite similar number of points in all classes. The next inquiry is made for comparing the results point by point on each method.

	Number of points	Ratio
Same class	13724	88%
Two different classes	1883	12%
Three different classes	10	less than 1%

Table 3: Number of classes on point base

88% of points resulted in the same class by all methods, the other 12% of them had two different classes and 10 points had three different classes (Table 3). A major source of differences is the side of the street, where grass and a sidewalk are present (Figure 12). The crucial points are in the class of grass (711) and in the pavement (453). This suggests that the problematic area is around the sidewalk. The other two classes have less than 20 problematic points.

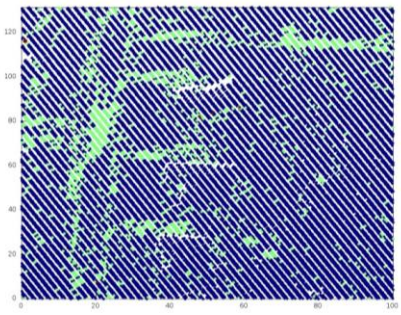


Figure 12: Number of classes assigned to each point (blue: same class for all methods, green: two different classes)

5.3 Validation

The validation is based on a manually classified dataset. The results were compared to an area that includes 910 points. The results of four methods are very similar with and without the additional parameters. As described above, the difficult part of the classification is the area around the sidewalk. In this area, the two classification methods have some differences; SOM classifies the whole area as pavement (even with different input parameters), while Bayes classification results in incorrect classification and it affects the ratio of point number per classes. In the view of numeric results, the SOM with fitting has the best performance with 95% of correctly classified points, and the wavelets with Bayes classification gave the worst with 91%.

Fitting and SOM	95.05%
Fitting and Bayes	92.64%
Wavelet and SOM	91.32%
Wavelet and Bayes	90.99%

Table 4: Validation results

6. CONCLUSION

The goal of the investigation was to find out if classification based solely on waveform data is feasible in urban areas. For this purpose different methods were tested and also special parameters were added. All four methods gave reasonable results but the common used parameters (intensity, range differences, number of returns) can't be overlooked. The pavement and roof can be separated well by all of the methods. For trees, however, the multi-echo based classification is needed and the proper separation of roof and pavement requires range difference (or height difference) calculation.

There were no significant differences between the generalized Gaussian fitting and wavelet transformation derived parameters in terms of classification performance. SOM and Bayes classifiers showed significant differences in the sidewalk areas.

In summary, waveform data can be used for classification purposes over an urban region, but it does not always provide a consistent performance. The incidence angle has high impact on the shape of the waveform signal, and, as in the case of roofs, this depends on the slope direction of the roof and the actual flight direction and scan angle. In this case, the classification based on waveforms has lower accuracy. The wide range of roofing materials has also a negative effect on the accuracy of the classification.

ACKOWLEDGEMENT

The authors thank Woolpert Incorporated for the data provided for this research.

REFERENCES

Chauve, A., Mallet C., Bretar, F., Durrieu, S., Deseilligny, M. P., Puech, W., 2007. Processing Full-Waveform LiDAR Data: Modelling Rawsignals, ISPRS Workshop on Laser Scanning 2007 and SilviLaser 2007, Espoo, September 12-14, 2007, Finland, pp 102-107.

Buckheit, J. B., Donoho, D. L., 1995. Wavelab and reproducible research. Wavelets and Statistics, Springer-Verlag.

Duong, H., Pfeifer N., Lindenbergh, R., 2006. Full wave form analysis: ice sat laser data for land coverclassification Proceedings of ISPRS commission VII symposium: remote sensing: from pixels to processes, Enshede, Netherlands, pp 185-195.

Kohonen, T., 1990. The Self-Organizing Map. Proc. IEEE 78(9), pp. 1464–1480.

Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J., 1996. Som pak, the Self-Organizing Map Program Package. Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, FIN-02150 Espoo, Finland, 27 pages.

Laky, S., Zaletnyik, P., Toth, C., 2010. Compressing LiDAR waveformdata. Proceedings of the International LiDAR Mapping Forum 2010, Denver, pp 1-10.

P.J. Green, 1995. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination, *Biometrika*, vol. 82, pp. 711-732.

Zaletnyik, P., Laky, S., Toth, C., 2010. LiDAR Waveform Classification Using Self-Organizing Map. Proceedings of the ASPRS 2010 Annual Conference, San Diego, USA, pp 1-12.

VEHICLE DETECTION FROM AN IMAGE SEQUENCE COLLECTED BY A HOVERING HELICOPTER

Fatemeh Karimi Nejadasl[a] and Roderik C. Lindenbergh[b]

[a]Department of Molecular Cell Biology, Leiden University Medical Center
P.O. Box 9600, 2300RC Leiden, The Netherlands, fkariminejadasl@lumc.nl
[b]Dept. of Remote Sensing, Delft University of Technology
Kluyverweg 1, 2629 HS, The Netherlands, r.c.lindenbergh@tudelft.nl

Working Groups I/2, III/1, III/4, III/5

KEY WORDS: Foreground and background identification, vehicle detection, temporal profile, sequential method, maximum frequency

ABSTRACT:

This paper addresses the problem of vehicle detection from an image sequence in difficult cases. Difficulties are notably caused by relatively small vehicles, vehicles that appear with low contrast or vehicles that drive at low speed. The image sequence considered here is recorded by a hovering helicopter and was stabilized prior to the vehicle detection step considered here. A practical algorithm is designed and implemented for this purpose of vehicle detection. Each pixel is identified firstly as either a background (road) or a foreground (vehicle) pixel by analyzing its gray-level temporal profile in a sequential way. Secondly, a vehicle is identified as a cluster of foreground pixels. The results of this new method are demonstrated on a test image-sequence featuring very congested traffic but also smoothly flowing traffic. It is shown that for both traffic situations the method is able to successfully detect low contrast, small size and low speed vehicles.

1 INTRODUCTION AND TEST DATA DESCRIPTION

Traffic is a problem of all large cities and is continuously analyzed by both authorities and researchers. Driving behavior is the most influential element in traffic and still less is known about it. This is due to the lack of instruments to track many vehicles for a long period of time without their awareness in taking part in an experiment (Ossen, 2008).

For the purpose of studying driving behavior in real traffic situations (Ossen et al., 2006; Ossen, 2008), a freeway is observed by a camera mounted below a hovering helicopter. The helicopter flies in the range of 300 to 500 m above the freeway and records image sequences for a period of maximum half an hour. The camera used for this purpose has a black and white visual sensor, a frequency of 15 frames per second and a resolution of 1392×1040 pixels. An area of 300 to 500 meter on the ground was covered with a spatial resolution of 20 to 50 centimeter. In this paper, a test data set is considered of an image sequence with 1501 images. The image sequence is previously stabilized on the road area which is the region of interest in this research.

In this paper, vehicles and background are identified from the image sequence of the stabilized road area. The specific problems of the vehicle detection in our data set are caused by slow moving vehicles,vehicles that appear with low contrast and small vehicles with a low amount of detail.

A vehicle here is extracted as a cluster of pixels (blob). First pixels are divided into two groups, one group consists of background pixels and the other group corresponds to moving objects. The division is performed based on the temporal profile of each pixel. The foreground pixels are grouped as a blob.

The paper is organized as follows. In Section 2, a brief review of related literature is presented. In Section 3, the foreground and background identification method is sketched and results are presented in Section 4. Conclusions are drawn in Section 5.

2 RELATED WORK

Vehicles can be detected by model-based methods, a 2D or 3D shape or/and an intensity template for a vehicle. The objective is to find this template back in each image considered, (Ballard, 1981; Ferryman et al., 1995; Tan et al., 1998; Pece and Worrall, 2002; Hinz, 2003; Zhao and Nevatia, 2003; Dahlkamp et al., 2004; Pece, 2006; Ottlik and Nagel, 2008). The disadvantage of model-based methods is the high dependency on geometric details of the considered object, which in our case would require that vehicles should appear in the image sequence with many details and with clear boundaries. In our data set, the shape and appearance of cars are simple are lacking detail. The similarity of some of the vehicles to road stripes, moreover, may cause a failure for the model-based methods.

Stauffer and Grimson (1999) modeled the background PDF as a mixture of Gaussians: usually three to five Gaussian models are enough for a complex background with illumination variation and background movement such as swaying trees. The value of each pixel in time, intensity or color, is a mixture of Gaussians. The parameters of the mixture of Gaussians model are weight, mean and standard deviation of each Gaussian model. They are estimated in an adaptive way. For every new image, the new observation, a pixel value, only updates the Gaussian parameters it belongs to. If a new observation does not belong to any Gaussian model, it constructs a new Gaussian model. The last Gaussian model, the Gaussian with the smallest weight, is combined with the Gaussian model with the second smallest weight. As a result, this pixel is assumed to belong to a moving object. Each parameter of the Gaussian model is updated as a combination of its previous value and the value of the new pixel. The weight of the pixel value is set by a learning parameter. A larger learning value increases the chance of wrongly modeling the object as background. A smaller value, however, cannot be used for very fast changing backgrounds. After background pixel identification, object pixels are connected to reconstruct the blob. In Stauffer and Grimson (2000), blobs are tracked using a Kalman

filter with multiple models. Every time a pool of blobs is evaluated against a pool of models. The model that explains the blobs best is used as a tracking result.

Pece (2002) assumed a mixture of probability likelihood models for both background and object clusters. The probability model of a cluster is the multiplication of position and gray-level PDF models. The background position is assumed to have a uniform distribution. The background is subtracted from each image to construct the difference image. The gray levels of the difference-image follow a Laplacian distribution (two-sided exponential). The object model for the position has a Gaussian or a top-hat distribution and the gray level has a uniform distribution. Expectation maximization is used to estimate the parameters of the PDF models. Clusters are analyzed for merging, splitting or making a new cluster. The current position of each cluster initializes the position of the cluster in the next image. The expectation maximization calculates the new location of the clusters. The camera should be fixed.

In Elgammal et al. (2002), a PDF model of a pixel value, color or intensity, is modeled as a mixture of nonparametric kernel models. The kernel model assumes a Gaussian shape, where the mean is the value of the pixel in the previous image and the the variance is defined as $\frac{1}{0.68\sqrt{2}}$ of the median value of the pixel temporal profile. A set of a recent pixel values is used for estimation of the PDF model of the pixel value. The pixel is considered as an object pixel if the pixel probability is less than a predefined threshold. The pixels assigned as background are grouped as a region. The region is only considered as background if the maximum probability of the region is larger than a predefined threshold. Otherwise the region is classified as an object. Remaining falsely detected background pixels are removed by a test against another predefined threshold value based on the product of the probability values of pixels in the connected region: if this product is lower than the threshold, the region is considered an object. The images should be obtained without camera motion.

Inspired by the works of Stauffer and Grimson (1999); Pece (2002); Elgammal et al. (2002), we have designed and implemented a practical approach to identify vehicles in the data set that also consists of vehicles, which are either moving slow, or are small, or are having low contrast. Such vehicles are, in existing approaches, easily grouped as belonging to the background.

3 VEHICLE DETECTION

Differences between a specific image and consecutive images can be used to initialize the motion detection. The illumination variation is considered negligible between consecutive images. The difference images have values around zero for a background and values larger than zero for locations where moving objects are present. At those locations where vehicles overlap, usually values near zero occur as well. The region after and before the overlap area shows significant differences (Figure 1).

Non-zero areas in a difference image will be relatively small in case either slow vehicles or low contrast vehicles are present. As a consequence, a low contrast, small sized or slow vehicle will be easily discarded as noise.

The solution to the slow motion problem is to use a background image, an image without any vehicles, instead of consecutive images to highlight the vehicles (Figure 2). The problem however is how to construct such background image considering illumination variation.

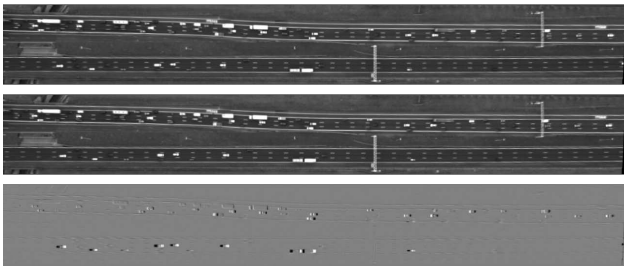


Figure 1: The difference image (bottom) is obtained by subtracting the middle from the top image. The top and middle images are consecutive.

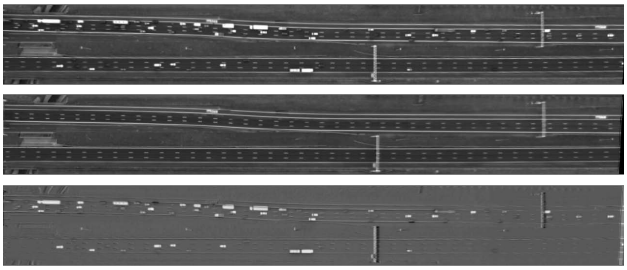


Figure 2: The difference image (bottom) is obtained by subtracting the background image (middle image) from the top image.

For every pixel, a time series of gray level observations can be constructed by analyzing the image sequence, which is called a temporal profile. The histogram of gray levels in each profile shows a distinct peak for the gray level of the background. The shape of the histogram is represented by a single Gaussian when the illumination variation is gradual and when the most of the time the pixel is not occupied by vehicles. The tail of the Gaussian shows the gray value of the foreground (Figure 3).

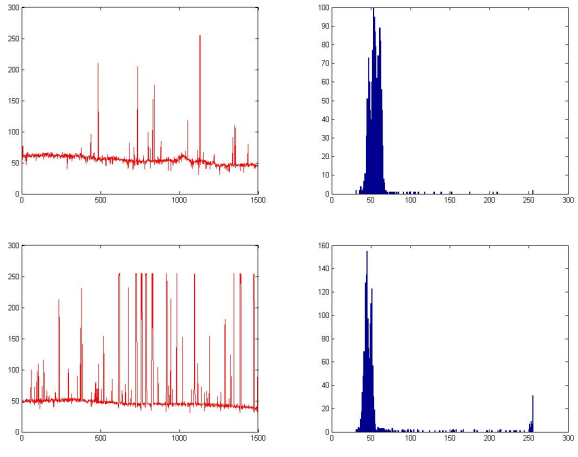


Figure 3: Top: the left graph is the temporal profile of a selected pixel gray level as a function of time, that is, the gray level (y-axis) is given as a function of the image number in the image sequences (x-axis). The top right graph shows the histogram of gray levels of the same profile. Bottom: similarly for another pixel.

The distribution of the gray levels in the profile of a particular pixel, therefore, identifies which pixel value belongs to the background and which one to the foreground. The most reliable way to determine the frequency distribution of the gray levels is by

processing all available observations, i.e. all images including a particular pixel. This operation requires a relatively long time however.

We have developed an efficient and fast way to identify background pixels on the basis of observed gray level frequencies without processing all available observations. This procedure is based on a sequential search on the gray level temporal profile for each pixel.

All pixels in the first image are assigned to the background class. Next, for each pixel, the gray level in the next image is compared to the background value. If the value is within ϵ gray levels of the background gray-level, the pixel is classified as background and the background value is updated to the value of this pixel. Here $\epsilon = \pm 6$ is used for an image gray-level range of $0 - 255$. The frequency of the background is updated by one. If the value falls outside an ϵ gray level interval, the pixel is classified as a foreground (1st identified object). Then the following image is processed, i.e. the following observation for any given pixel is analyzed in the same way as in the previous step.

The comparison with the current background gray level and statistics is done exactly as in the previous step. If the new observation falls within a +/- 6 gray level interval of the value associated with the 1st identified object, the pixel value and its probability are updated in the same way as for the background. If it falls outside this range a new object with a frequency of one is created (last identified object). The following observation is compared with the background and last identified object gray-levels. When a gray level is observed which falls within the ϵ range of either background or last identified object, the corresponding gray-level and frequency are updated. If the gray-level falls outside both ranges, a new object is identified with a frequency of one. The two previously identified objects are retained but not used for the analysis of subsequent observations. When a new observation falls outside the gray level ranges associated with the background and the last identified object, the oldest object in the list is removed.

The gray level of the last identified object is also compared with the gray level of its predecessor. If it falls inside the range associated with the latter, the frequencies are added and the predecessor is removed. The frequencies of the current background and last identified objects are compared and the larger frequency is assigned to the background class and the lower one to the last identified object. The corresponding gray levels are exchanged as well. This procedure corrects the initial assignment of all pixels to the background class. Moreover, it prevents erroneously assigned background gray levels from propagating to other images. Likewise, the procedure improves progressively the estimate of the last-identified-object gray level. Figure 4 shows the result of the background/foreground pixel identification.

The details of this algorithm is described in Algorithm 1.

The foreground pixels are connected by a 8-neighborhood connected component to cluster a group of foreground pixels, called a blob. The blob here represents the vehicle. Very small blobs are assumed to be caused by noise and are removed by morphological opening with a 3×3 structural element (Gonzalez and Woods, 2007). Figure 5 represents the extracted blobs.

4 RESULTS AND DISCUSSIONS

Algorithm 1 is applied on the test data consisting of 1501 images. Figure 6 shows the identified vehicles in yellow. Two regions

Algorithm 1: Near real-time foreground and background pixel identification. V and P represent respectively a value and a probability. The subscripts b, f, $f1$, and $f2$ denote background and three reference objects respectively. $\mathbf{0}$ and $\mathbf{1}$ are matrices with all elements equal to zero and one respectively. The subscript one for image I and classified image BF indicates the first image.

Input: Stabilized image-sequence of the road area ($\{I\}$)
initialized by $BF_1 \leftarrow \mathbf{1}, V_b \leftarrow I_1, P_b \leftarrow \mathbf{1}$,
$V_f, P_f, V_{f1}, P_{f1}, V_{f2}, P_{f2} \leftarrow \mathbf{0}$

1 **for** f = *image number* **do**
2 **for** y = *row* **do**
3 **for** x = *column* **do**
4 **if** $I(x,y) \in [V_b(x,y) \pm \epsilon]$ **then**
5 $V_b(x,y) = I(x,y,f)$
6 $P_b(x,y) = P_b(x,y) + 1$
7 $BF(x,y,f) = 1$
8 **else**
9 **if** $I(x,y,f) \in [V_f(x,y) \pm \epsilon]$ **then**
10 $V_f(x,y) = I(x,y,f)$
11 $P_f(x,y) = P_f(x,y) + 1$
12 $BF(x,y,f) = 0$
13 **else**
14 $V_{f1}(x,y),\ P_{f1}(x,y) \rightarrow$
$V_{f2}(x,y),\ P_{f2}(x,y)$
15 $V_f(x,y),\ P_F(x,y) \rightarrow$
$V_{f1}(x,y),\ P_{f1}(x,y)$
16 $V_f(x,y) = I(x,y,t),\ P_f(x,y) = 1$
17 $BF(x,y,f) = 0$
18 **if** $V_f(x,y) \in [V_{f2}(x,y) \pm \epsilon]$ **then**
19 $P_f(x,y) = P_{f2}(x,y) + P_f(x,y)$
20 $V_{f2}(x,y) = 0,\ P_{f2}(x,y) = 0$
21 **if** $P_b(x,y) < P_f(x,y)$ **then**
22 $V_f(x,y),\ P_f(x,y) \leftrightarrow$
$V_b(x,y),\ P_b(x,y)$
23 $BF(x,y,f) = 1$

Output: New image-sequence with background and foreground identified($\{BF\}$)

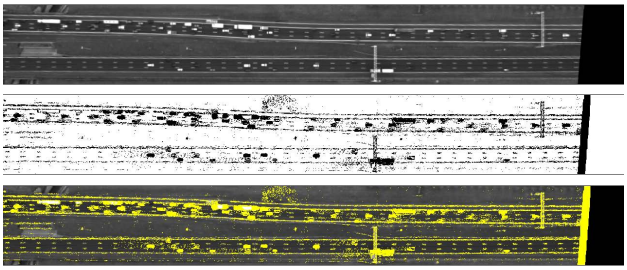


Figure 4: Background/foreground pixel identification. From top to bottom: 1) an arbitrary image, 2) discrimination of the background (white) and foreground (black) pixels, and 3) the detected foreground pixels highlighted.

from the upper and lower part of the road are considered in the zoom-in.

The upper lanes in the road image shows congested traffic while the lower lanes have fluid traffic. To evaluate the performance of the identification method on these two different type of traffic, seven images are selected and the results are displayed in

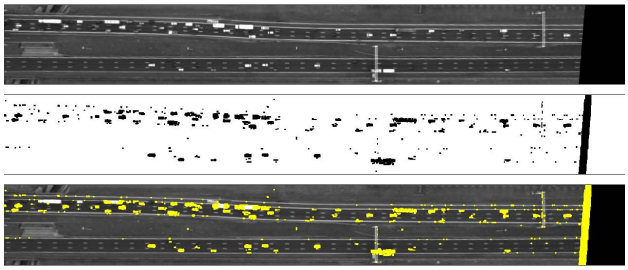


Figure 5: Vehicle identification. From top to bottom: 1) an arbitrary image 2) extracted blobs (black), and 3) highlighted extracted blobs.

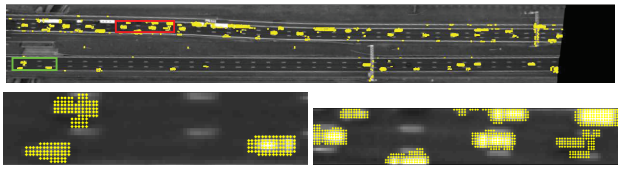


Figure 6: Identified vehicles in the image 100 (top), the zoomed area depicted by a green rectangle (bottom left), the zoomed area depicted by a red rectangle (bottom right).

Figure 7. The statistics come in Table 1 for fluid traffic and in Table 2 for congested traffic. All validation is based on careful visual inspection of the results.

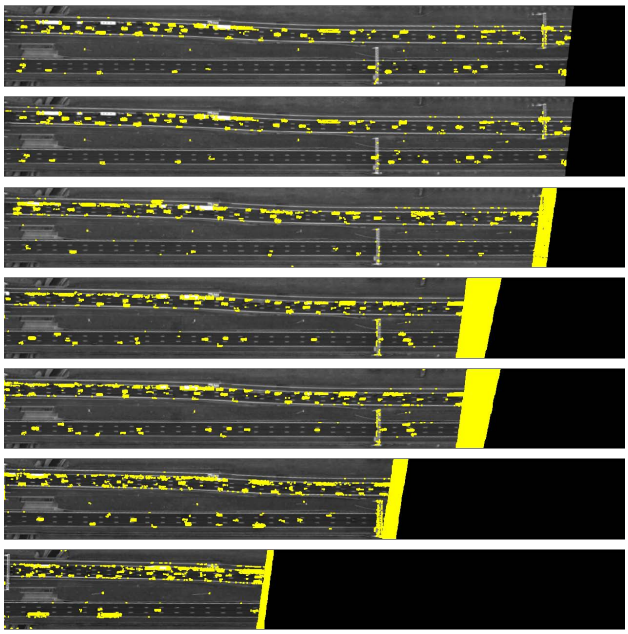


Figure 7: Identified vehicles: top to bottom: image number 100, 101, 248, 500, 501, 1000 and 1500.

In each table, the total number of vehicles, the number of identified vehicles, missing vehicles, wrongly identified vehicles, mixed vehicles and vehicles which are identified as more than one vehicle are listed. Only wrongly detected vehicles on the road area are counted, vehicles found by the algorithm on road lines and a gantry were discarded.

Vehicles which were missed in one image can be identified in an other image. For example in images 101 and 501 one of the vehicles which were not detected in the images 100 and 500 respectively, were identified. In image 101, one vehicle leaves the area. Therefore the total number of correctly detected vehicles is the same for both images 100 and 101.

Table 1: Identification results for smoothly flowing traffic. N, T, TP, FN, Mx and Dj are respectively the image number, the total number of vehicles, the number of true positives (correctly identified vehicles), false negatives (not identified vehicles), false positives (wrongly identified vehicles), mixed vehicles and disjoint regions.

N	T	TP	FN	FP	Mx	Dj
100	18	17	1	0	0	1
101	17	17	0	0	0	2
248	9	9	0	1	0	1
500	19	17	2	1	1	2
501	19	18	1	0	0	0
1000	17	15	2	1	1	0
1500	6	6	0	0	0	1

Table 2: Identification results for very congested traffic. The labels are as identified in Table 1.

N	T	TP	FN	FP	Mx	Dj
100	59	50	9	3	4	4
101	59	50	9	1	4	1
248	64	59	5	3	20	5
500	68	57	11	1	9	4
501	68	57	11	1	6	3
1000	52	47	5	6	3	3
1500	35	30	5	10	2	3

The vehicle identification method works well in case of moving traffic. The results is however less good in the case of congested traffic. The main problem in this case is the mixing of vehicles which sometimes happens when vehicles are too close. This problem cannot be solved until the vehicles start to separate. Another problem is less serious: sometimes one vehicle in reality leads to more than one detected vehicle. This problem occurs mainly on trucks with very slow movement. The number of disjoint vehicles is relatively low however.

5 CONCLUSIONS

In this paper, vehicles were detected from the image-sequence stabilized on the road area as recorded by a hovering helicopter. It turns out to be possible to extract vehicles, as blobs, in several difficult situations. Previous problems with the vehicle detection of i) small size vehicles observed with scarce detail, ii) vehicles with low speed, and iii) low contrast vehicles, could be largely solved. A very low amount of memory is needed for the processing. Because each pixel is only compared to the previous value of the background, the procedure is done in a sequential way for every image. Besides, the procedure starts without any extra assumptions and more importantly, there is a mechanism to recover the background when a pixel is wrongly classified as a vehicle. Moreover, the background value is always updated to the current gray level. As a result, this method also works for image sequences representing heavy traffic conditions and for image sequences that are exposed to gradual illumination variations. By looking at a temporal profile of a pixel, it has been observed that illumination variations of this pixel gray-level for the background

could be large, but these background illumination variations are negligible if just a few consecutive images are considered. A sudden large illumination variation, however, cannot be handled by this method.

Although in general only part of a very low contrast vehicle can be extracted by this method, tracking of such a vehicle using our algorithm is still possible, compare (Karimi Nejadasl et al., 2006). Vehicles, which are too close to each other, are grouped as one blob, thus their tracking is unreliable. When these vehicles start to take a distance from each other, they are identified as separate blobs and then tracked reliably. However, if a vehicle stays very long in one place it is classified as background. When the vehicle starts to move, updating the background value requires more images to be processed.

ACKNOWLEDGEMENTS

The research presented in this paper is part of the research program "Tracing Congestion Dynamics with Innovative Traffic Data to a better Theory", sponsored by the Dutch Foundation of Scientific Research MaGW-NWO. We like to thank Ben Gorte, Massimo Menenti and Serge Hoogendoorn for helpful discussions.

References

Ballard, D. H., 1981. Generalizing the hough transform to detect arbitrary shapes. Pattern Recognition 13(2), pp. 111–122.

Dahlkamp, H., Pece, A. E., Ottlik, A. and Nagel, H. H., 2004. Differential analysis of two model-based vehicle tracking approaches. DAGM LNCS 3175, pp. 7178.

Elgammal, A., Duraiswami, R., Harwood, D. and Davis, L. S., 2002. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. Proceedings of the IEEE 90(7), pp. 1151–1163.

Ferryman, J. M., Worrall, A. D., Sullivan, G. D. and Baker, K. D., 1995. A generic deformable model for vehicle recognition. BMVC95 1, pp. 127–136.

Gonzalez, R. C. and Woods, R. E., 2007. Digital Image Processig. 3 edition edn, Prentice Hall.

Hinz, S., 2003. Integrating local and global features for vehicle detection in high resolution aerial imagery. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 34(Part 3/W8), pp. 119–124.

Karimi Nejadasl, F., Gorte, B. G. H. and Hoogendoorn, S. P., 2006. Optical flow based vehicle tracking strengthened by statistical decisions. ISPRS Journal of Photogrammetry and Remote Sensing 61(3-4), pp. 159–169.

Ossen, S. J. L., 2008. Longitudinal driving behavior: theory and empirics. PhD thesis.

Ossen, S. J. L., Hoogendoorn, S. P. and Gorte, B. G. H., 2006. Interdriver differences in car-following: a vehicle trajectory-based study. Transportation Research Record 1965, pp. 121–129.

Ottlik, A. and Nagel, H. H., 2008. Initialization of model-based vehicle tracking in video sequences of inner-city intersections. International Journal of Computer Vision.

Pece, A. E., 2006. Contour tracking based on marginalized likelihood ratios. Image and Vision Computing 24, pp. 301–317.

Pece, A. E. and Worrall, A. D., 2002. Tracking with the em contour algorithm. European Conference Computer Vision pp. 3–17.

Pece, A. E. C., 2002. Generative-model-based tracking by cluster analysis of image differences. Robotics and Autonomous Systems 39(3-4), pp. 181–194.

Stauffer, C. and Grimson, W., 1999. Adaptive background mixture models for real-time tracking. IEEE Int'l Conf. on Computer Vision and Pattern Recognition.

Stauffer, C. and Grimson, W. E. L., 2000. Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Tan, T., Sullivan, G. and Baker, K., 1998. Model-based localisation and recognition of road vehicles. International Journal of Computer Vision 27(1), pp. 5–25.

Zhao, T. and Nevatia, R., 2003. Car detection in low resolution aerial images. Image and Vision Computing 21(8), pp. 693–703.

MOTION COMPONENT SUPPORTED BOOSTED CLASSIFIER FOR CAR DETECTION IN AERIAL IMAGERY

Sebastian Tuermer[a], Jens Leitloff[a], Peter Reinartz[a], Uwe Stilla[b]

[a] Remote Sensing Technology Institute, German Aerospace Center (DLR)
Oberpfaffenhofen, Germany
sebastian.tuermer@dlr.de, jens.leitloff@dlr.de, peter.reinartz@dlr.de,
[b] Photogrammetry and Remote Sensing, Technische Universitaet Muenchen (TUM)
Arcisstrasse 21, 80333 Munich, Germany
stilla@tum.de

Commission III/5

KEY WORDS: Vehicle detection, AdaBoost, HoG features, Aerial image sequence, Motion mask

ABSTRACT:

Research of automatic vehicle detection in aerial images has been done with a lot of innovation and constantly rising success for years. However information was mostly taken from a single image only. Our aim is using the additional information which is offered by the temporal component, precisely the difference of the previous and the consecutive image. On closer viewing the moving objects are mainly vehicles and therefore we provide a method which is able to limit the search space of the detector to changed areas. The actual detector is generated of HoG features which are composed and linearly weighted by AdaBoost. Finally the method is tested on a motorway section including an exit and congested traffic near Munich, Germany.

1 INTRODUCTION

Already within the last century the impact and the significance of mobility and especially individual traffic has increased enormously (Banister et al., 2010). The phenomenon results in overloaded streets and highways. Further this leads to environmental pollution, wast of resources and finally threatens humans' quality of life (Ouis, 2001).
To adequately overcome this problem, scientists worldwide are working on smart solutions. They all need data of realistic traffic scenarios which can be analyzed and evaluated. Final goal are strategies to improve the current traffic situation. Mainly two applications should be named in the real-time case, mass events and catastrophes. Manager of mass events will be able to canalize the usual high volume of traffic. This results in a higher security level. Also emergency teams and rescue crews are supported by traffic data in the event of a disaster. They will be able to choose the fastest ways reaching the affected area and can see in detail where to set up a control room or a collection point. Due to these important applications there are some other procedures of gathering traffic information besides the optical ones. For instance induction loops, light barriers, radar based methods or floating car solutions. But all of these methods are not suitable for monitoring a wide area consistently.
We present a method for extracting vehicles in sequential aerial imagery. The method uses HoG features and Boosting as machine learning algorithm. The focus lies on the motion mask which affords detection of moving objects faster and more reliable.

2 RELATED WORK

Methods for vehicle detection in optical images often belong to one of three groups according to the platform of the sensor. The field with definitely the highest amount of research activity during the last years are stationary video cameras which provide side view images or at least oblique view images. Further property is a quite high imaging frequency in comparison to the other groups. The use of wavelet coefficients as features and AdaBoost can be seen in (Schneiderman and Kanade, 2000). Also (She et al., 2004) are detecting cars by the use of Haar wavelets features in the HSV color space. A combination of Haar and HoG features which are formed to a strong cascading classifier by Boosting presents (Negri et al., 2008). In (Kasturi et al., 2009) a simple background subtraction is done which is only working for video data. An overview on the work for stationary cameras can be found in (Sun et al., 2006).
The next group considers satellite imagery which provide a reduced spatial resolution (highest resolution is often max 0.5 m) and mainly use single images, not time series. An approach which uses simple features based on shape and intensity presents (Eikvil et al., 2009). Using segmented images and applying a maximum likelihood classification can be observed in (Larsen et al., 2009). Promising results have also been achieved by (Leitloff et al., 2010). They use Haar-like features in combination with AdaBoost.
The last group of approaches deals with airborne images. At this step we first suggest a further separation in explicit or implicit models. Approaches based on explicit models are for example given in (Moon et al., 2002) with a convolution of a rectangular mask and the original image. Also (Zhao and Nevatia, 2003) offer an interesting method by creating a wire-frame model and try to match it with extracted edges at the end of a Bayesian network. A similar way is suggested by (Hinz, 2003a) (Hinz, 2003b), the author makes the approach more mature and added additional parameters like the position of the sun. (Kozempel and Reulke, 2009) provide a very fast solution which takes four special shaped edge filters trying to represent an average car. Another approach of (Reilly et al., 2010) shows a method which is based on background subtraction. The background is computed by a 10 frame median image.
Finally implicit modeling is used by (Grabner et al., 2008), they take Haar-like features, HoG features and LBP (local binary patterns). All these features are passed to an on-line AdaBoost training algorithm which creates a strong classifier.
Another approach using aerial data and trying to have benefit of

the temporal component, similar to our idea, is (Benedek et al., 2009). Their aim is not only the detection of cars but all moving objects. To realize this idea a three layer Markov random field model is introduced.
A comprehensive overview and evaluation of airborne sensors for traffic estimation can be found in (Hinz et al., 2006) and (Stilla et al., 2004).

3 METHOD

In general, the method is developed for airborne, high resolution frame camera systems with high imaging frequency. The workflow of our method is shown in Fig. 1. Following subsections give explanations to parts of the workflow or refer to related literature for detailed information.

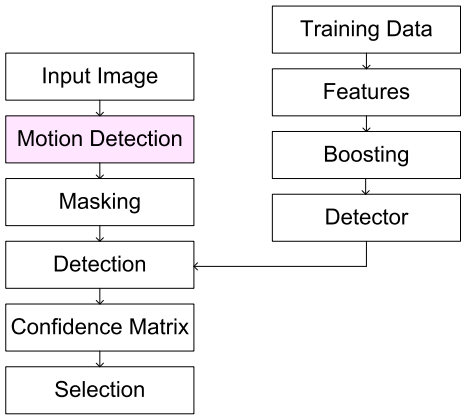


Figure 1: Workflow of proposed car detection method

3.1 Color Space

For our purpose we decided to use a color space which is technically oriented. That means per definition the color space is a linear transformation of the RGB color space. The utilized color space is named I1I2I3 and meets, according to (Ohta et al., 1980) and own tests, the requirements of the proposed method (Sec. 3.2) very well. Which is mainly the quality of the resulting difference image. Mathematically expressed the transformation is shown in Eq. 1:

$$\begin{bmatrix} I1 \\ I2 \\ I3 \end{bmatrix} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & -1/2 \\ -1/4 & 1/2 & -1/4 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

where R, G, B are the red, green, blue channels and I1, I2, I3 are the resulting channels of I1I2I3 color space model.

3.2 Motion Detection

The idea of the motion mask is based on turning all available information to account which is delivered by our camera system. To reach that aim a usual way of motion detection is processing a difference image. A difference image shows all pixels which have changed in comparison to the other image. One possibility is to calculate the difference image with the current image and its background image. Unfortunately the problem is that we do not have an image without foreground objects.
A solution of this problem offers the use of three images and a subtraction of each (Dubuisson and Jain, 1995). In detail, we calculate the difference of the current image and the previous image, and the difference of the current image and the subsequent image as well. The two resulting difference images are linked with the Boolean AND. The approach expressed in formulas can be seen in Eq. 2 where the first difference image D_1 is calculated (Rehrmann and Birkhoff, 1995):

$$D_1(t_1, t_2, x, y) =$$

$$\begin{cases} 1, \text{ if } & |I_{I1}(t_2, x, y) - I_{I1}(t_1, x, y)| \\ & + |I_{I2}(t_2, x, y) - I_{I2}(t_1, x, y)| \\ & + |I_{I3}(t_2, x, y) - I_{I3}(t_1, x, y)| > d_{min} \\ 0, \text{ else} \end{cases} \quad (2)$$

where the functions of the images are $I_{I1}(t, x, y)$, $I_{I2}(t, x, y)$ and $I_{I3}(t, x, y)$. The parameter t is a discreet time whereas x and y are the position in the image for the three different channels $I1$, $I2$, $I3$ of the color space. The parameter d_{min} is a threshold which is necessary for excluding intensity changes of pixels due to camera noise, various illuminations or the different illustration geometry.
Subject to the condition that we have 3 consecutive images the next step is linking the two difference images which is depicted in Eq. 3:

$$D_2(t_1, t_2, t_3, x, y) =$$

$$\begin{cases} 1, \text{ if } & D_1(t_1, t_2, x, y) = 1 \\ & \wedge D_1(t_2, t_3, x, y) = 1 \\ 0, \text{ else} \end{cases} \quad (3)$$

with $D_1(t_1, t_2, x, y)$ difference image of previous and current image and $D_1(t_2, t_3, x, y)$ difference image of current and consecutive image.

3.3 Features

We use HoG features (Dalal and Triggs, 2005) to differentiate cars from other objects. A reason for this choice is a test where Haar and HoG features are compared with regard to their car detection capability (Tuermer et al., 2011). HoG features are created by quantize gradient magnitudes to a histogram. The particular bin is chosen according to the gradient orientation. A detailed explanation of these features and how the feature extraction works can be found in (Tuermer et al., 2010).

3.4 Training

The training creates the custom classifier. We pass the extracted features of more than 400 car samples to the machine learning algorithm. This algorithm is part of the Boosting group (Freund and Schapire, 1997) (Freund and Schapire, 1999) and is named Real AdaBoost. Boosting is a method which builds a strong classifier by a weighted linear combination of weak classifiers. In our case a weak classifier is a threshold applied on a feature which is able to classify more accurate than 50 percent object of interest or not object of interest. The procedure of weighting and re-weighting is graphically explained in Fig. 2. The formula of the composite strong classifier H can be expressed as Eq. 4 shows:

$$H(X) = \boldsymbol{sign}(a_1 h_1(x) + a_2 h_2(x) + a_3 h_3(x)) \quad (4)$$

where a_i are weightings and h_i are weak classifiers.

3.5 Detection

The ordinary detection is done by sliding the previously generated classifier over the whole search image and applying it at every pixel position. A method which is time consuming and susceptible to mistakes. Alternatively, the proposed innovative

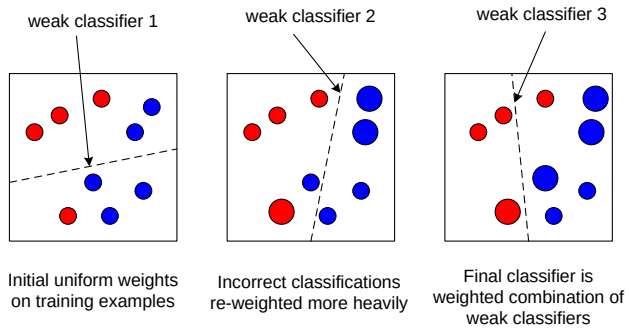


Figure 2: Boosting Schema

method just applies the detector where the motion mask is true. An additional graphical explanation can be found in Fig. 3. The response obtained from the classifier is a confidence value which has information how reliable the detection candidate is. Sometimes applying a threshold to the confidence matrix is necessary to adjust the result to the respective requirement. On the one hand it could be useful to detect all cars in the image and accept false positives as consequence. On the other hand it could be necessary to obtain correct detections only and accept false negatives.

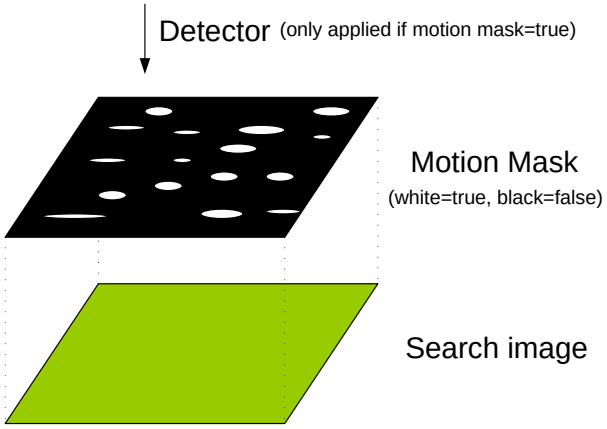


Figure 3: Functional description of the motion mask

4 CAMERA SYSTEM

The utilized aerial test data are acquired from the 3K camera system, which is composed of three off-the-shelf professional SLR digital cameras (Canon EOS 1Ds Mark II). These cameras are mounted on a platform which is specially constructed for this purpose. A picture of the cameras and the platform is shown in Fig. 4. Furthermore a calibration was done (Kurz et al., 2007) to enable the georeferencing process which is supported by GPS (Global Positioning System) and INS (inertial navigation system). The system is designed to deliver images with maximum 3 Hz recording frequency combined into one burst. A burst consists of 2 to 4 images and is necessary because otherwise the camera would not be able to write the data to the memory card. After one burst a pause of 10 seconds follows. Depending on the flight altitude a spatial resolution up to 15 centimeters (at 1000m altitude) is provided. For further information about the 3K camera system please refer to (Reinartz et al., 2010).

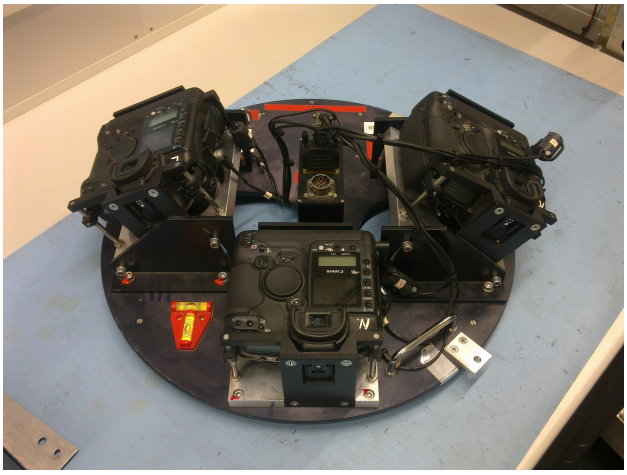


Figure 4: 3K camera system

5 EXPERIMENTAL RESULTS

The experimental results are based on image samples from a motorway in the east of Munich, Germany. Our intention is the detection of cars in two directions only (from right to left and vice versa); note the cars which take the exit have different orientations and are not classified. The search image (Fig. 5) is the second image out of three and thus imaged at time t_2 according to the preceding remarks (Sec. **??**). To give an impression how helpful the motion mask is, we display the result of a classification without motion mask in Fig. 6.
The next images show the genesis of the motion mask. The result of applying Eq. 2 can be seen in Fig. 7 and Fig. 8. The manual chosen threshold d_{min} amounts 30. But if necessary it can be easily substituted by the automatic Otsu thresholding method (Otsu, 1979). Applying Eq. 3 results in the final motion mask shown in Fig. 9. The remaining search space after applying the mask is depicted in Tab. 1. Finally the result of the proposed detection method is shown in Fig. 10. Where detections of moving vehicles are marked with red rectangles.

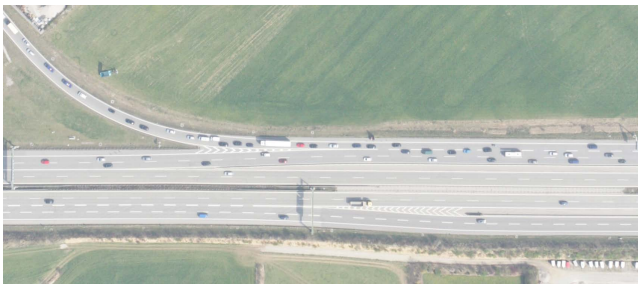


Figure 5: Original 3K image sample

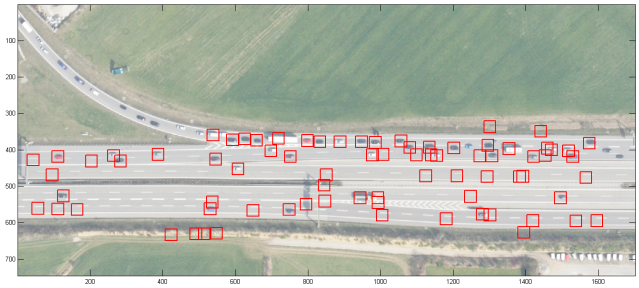


Figure 6: Classification result without motion mask

Figure 7: Difference image of image t_0 and t_1

Figure 8: Difference image of image t_1 and t_2

Figure 9: Boolean AND of the two difference images (Fig. 7, Fig. 8)

Figure 10: Classification result with motion mask

6 DISCUSSION

The car detection quality with motion mask (Fig. 10) is considerably enhanced compared to the test without motion mask (Fig. 6). This is due to the limited search space where static areas which include mainly no vehicles are excluded. But a consequence of this method is that also cars without or with low velocity are excluded. Now it might be necessary to develop a method which brings the detection methods for static and for moving cars together.
Concerning the detector design, it should be mentioned that there is still room for improvement as fas as the training data is concerned. We can trace the false positives in Fig. 6 back to the fact that the negative training sample database is not sufficient. These false car candidates often look very similar and are very often parts of the road with a small part of road markings. Perhaps it is possible to exclude them by a more intelligent training.
The advantage of the motion mask is not only the improved detection quality, but of course reduction of calculation time as well. A quick look at Tab. 1 shows that in the end only about one percent of the original test image have to be examined. This does not mean that the detector is 100 times faster, because it is a cascading detector and only the application of the first hierarchical level can be spared for all pixel positions. But calculating the motion mask is still faster than calculating all the features of the first hierarchical level of the detector.
Another interesting point in the processing chain of the motion mask itself is that the result in Fig. 8 has obviously much more disturbances than Fig. 7. This can be explained due to a lack of co-registration. The overlay of the images is only done by the use of the geocode and the relative error (image to image) of the georeferencing comes into full account. However the presented method is able to handle these kind of errors dependable. By the way the same result using RGB color space is much more noisy in comparison to the utilized I1I2I3 color space.

Table 1: Limited search space due to motion mask

	remaining search space of original image
$D_1(t_1, t_2, x, y)$ (Fig.7)	2.05 %
$D_1(t_2, t_3, x, y)$ (Fig.8)	6.03 %
$D_2(t_1, t_2, t_3, x, y)$ (Fig.9)	1.01 %

7 CONCLUSIONS AND FUTURE WORK

We present a vehicle detection method which is improved by using additional information provided by the temporal component. Making use of three consecutive images allows to determine the position of a moving car very accurately. The resulting mask shows potential to identify moving objects, which will help to make vehicle detection more reliably in the future. But there is also a catch to progress in the case of slowly moving vehicles. It can be observed that slowly moving vehicles with intent to take the exit of the highway are not captured perfectly. The same applies to non-moving objects. This happens because some pixels still have the same color as the pixels at t_{i-1}. In this case the method needs further development. Benefit of the proposed detection method for moving vehicles is:

- detection runs much faster (up to 37x)
- more robust and reliable
- very high detection quality

Running the tests with a more intelligent training and a extended training database is one point of future work. Furthermore we would like to use test images from more difficult areas near city centers for instance. And finally the detector itself will get an upgrade regarding the ability of being rotation invariant.
Of course the detection can be remarkable improved by using additional information that is not used till this day. The database with positive samples consists only of images that show a car. One idea is to not only us sample chips with just a single car inside, but introduce a training database which regards to the surrounding of the cars. This could be helpful to distinguish if an object is situated on the road or on a roof for for example.

REFERENCES

Banister, D., Browne, M. and Givonia, M., 2010. Transport reviews - the 30th anniversary of the journal. Transport Reviews: A Transnational Transdisciplinary Journal 30, pp. 1–10.

Benedek, C., Sziranyi, T., Kato, Z. and Zerubia, J., 2009. Detection of object motion regions in aerial image pairs with a multilayer markovian model image processing. IEEE Transactions on Image Processing 18(10), pp. 2303 – 2315.

Dalal, N. and Triggs, B., 2005. Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, IEEE Computer Society, San Diego, CA, USA, pp. 886 – 893.

Dubuisson, M. P. and Jain, A. K., 1995. Contour extraction of moving objects in complex outdoor scenes. International Journal of Computer Vision 14(1), pp. 83–105.

Eikvil, L., Aurdal, L. and Koren, H., 2009. Classification-based vehicle detection in high-resolution satellite images. ISPRS Journal of Photogrammetry and Remote Sensing 64, pp. 65–72.

Freund, Y. and Schapire, R. E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), pp. 119–139.

Freund, Y. and Schapire, R. E., 1999. A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence 14(5), pp. 771–780.

Grabner, H., Nguyen, T. T., Gruber, B. and Bischof, H., 2008. On-line boosting-based car detection from aerial images. ISPRS Journal of Photogrammetry and Remote Sensing 63(3), pp. 382 – 396.

Hinz, S., 2003a. Detection and counting of cars in aerial images. In: International Conference on Image Processing (ICIP), Vol. 3, pp. 997–1000.

Hinz, S., 2003b. Integrating local and global features for vehicle detection in high resolution aerial imagery. In: Photogrammetric Image Analysis (PIA), Vol. 34(3/W8), International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, pp. 119–124.

Hinz, S., Bamler, R. and Stilla, U., 2006. Editorial theme issue: Airborne und spaceborne traffic monitoring. ISPRS Journal of Photogrammetry and Remote Sensing 61(3-4), pp. 135–136.

Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V. and Zhang, J., 2009. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2), pp. 319–336.

Kozempel, K. and Reulke, R., 2009. Fast vehicle detection and tracking in aerial image bursts. In: CMRT09, Vol. 38(3/W4), IAPRS, pp. 175–180.

Kurz, F., Mller, R., Stephani, M., Reinartz, P. and Schroeder, M., 2007. Calibration of a wide-angel digital camera system for near real time scenarios. In: ISPRS Hannover Workshop: High-Resolution Earth Imaging for Geospatial Information.

Larsen, S. O., Koren, H. and Solberg, R., 2009. Traffic monitoring using very high resolution satellite imagery. Photogrammetric Engineering and Remote Sensing 75(7), pp. 859–869.

Leitloff, J., Hinz, S. and Stilla, U., 2010. Vehicle extraction from very high resolution satellite images of city areas. IEEE Trans. on Geoscience and Remote Sensing 48, pp. 1–12.

Moon, H., Chellappa, R. and Rosenfeld, A., 2002. Performance analysis of a simple vehicle detection algorithm. Image and Vision Computing 20(1), pp. 1–13.

Negri, P., Clady, X., Hanif, S. M. and Prevost, L., 2008. A cascade of boosted generative and discriminative classifiers for vehicle detection. EURASIP Journal on Advances in Signal Processing 2008, pp. 1–12.

Ohta, Y.-I., Kanade, T. and Sakai, T., 1980. Color information for region segmentation. Computer Graphics and Image Processing 13, pp. 222–241.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Sys., Man., Cyber. 9(1), pp. 6266.

Ouis, D., 2001. Annoyance from road traffic noise: A review. Journal of Environmental Psychology 21(1), pp. 101–120.

Rehrmann, V. and Birkhoff, M., 1995. Echtzeitfhige Objektverfolgung in Farbbildern. In: Tagungsband 1. Workshop Farbbildverarbeitung, Fachberichte Informatik 15/95, University of Koblenz, pp. 36–39.

Reilly, V., Idrees, H. and Shah, M., 2010. Detection and tracking of large number of targets in wide area surveillance. In: European Conference on Computer Vision (ECCV) 2010.

Reinartz, P., Kurz, F., Rosenbaum, D., Leitloff, J. and Palubinskas, G., 2010. Near real time airborne monitoring system for disaster and traffic applications. In: Optronics in Defence and Security (Optro), Paris, France.

Schneiderman, H. and Kanade, T., 2000. A statistical method for 3d object detection applied to faces and cars. In: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, pp. 746–751.

She, K., Bebis, G., Gu, H. and Miller, R., 2004. Vehicle tracking using on-line fusion of color and shape features. In: International IEEE Conference on Intelligent Transportation Systems, pp. 731–736.

Stilla, U., Michaelsen, E., Soergel, U., Hinz, S. and Ender, J., 2004. Airborne monitoring of vehicle activity in urban areas. In: International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 34(Part B3), pp. 973–979.

Sun, Z., Bebis, G. and Miller, R., 2006. On-road vehicle detection: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(5), pp. 694–711.

Tuermer, S., Leitloff, J., Reinartz, P. and Stilla, U., 2010. Automatic vehicle detection in aerial image sequences of urban areas using 3d hog features. In: International Archives of Photogrammetry, Remote Sensing and the Spatial Information Sciences, Vol. XXXVIII(Part 3), Paris, France.

Tuermer, S., Leitloff, J., Reinartz, P. and Stilla, U., 2011. Evaluation of selected features for car detection in aerial images. In: Hanover Workshop 2011.

Zhao, T. and Nevatia, R., 2003. Car detection in low resolution aerial image. Image and Vision Computing 21(8), pp. 693–703.

AUTOMATIC CROWD ANALYSIS FROM VERY HIGH RESOLUTION SATELLITE IMAGES

Beril Sirmacek, Peter Reinartz

German Aerospace Center (DLR), Remote Sensing Technology Institute
PO Box 1116, 82230, Wessling, Germany
(Beril.Sirmacek, Peter.Reinartz)@dlr.de

Commission VII

KEY WORDS: Very high resolution satellite images, Crowd detection, DEM, Local features, Probability theory, Shadow extraction, Road extraction

ABSTRACT:

Recently automatic detection of people crowds from images became a very important research field, since it can provide crucial information especially for police departments and crisis management teams. Due to the importance of the topic, many researchers tried to solve this problem using street cameras. However, these cameras cannot be used to monitor very large outdoor public events. In order to bring a solution to the problem, herein we propose a novel approach to detect crowds automatically from remotely sensed images, and especially from very high resolution satellite images. To do so, we use a local feature based probabilistic framework. We extract local features from color components of the input image. In order to eliminate redundant local features coming from other objects in given scene, we apply a feature selection method. For feature selection purposes, we benefit from three different type of information; digital elevation model (DEM) of the region which is automatically generated using stereo satellite images, possible street segment which is obtained by segmentation, and shadow information. After eliminating redundant local features, remaining features are used to detect individual persons. Those local feature coordinates are also assumed as observations of the probability density function (pdf) of the crowds to be estimated. Using an adaptive kernel density estimation method, we estimate the corresponding pdf which gives us information about dense crowd and people locations. We test our algorithm using Worldview-2 satellite images over Cairo and Munich cities. Besides, we also provide test results on airborne images for comparison of the detection accuracy. Our experimental results indicate the possible usage of the proposed approach in real-life mass events.

1 INTRODUCTION

Recently automatic detection of people and crowds from images gained high importance, since it can provide very crucial information to police departments and crisis management teams. Especially, detection of very dense crowds might help to prevent possible accidents or unpleasant conditions to appear. Due to their limited coverage of area, street or indoor cameras are not sufficient for monitoring big events. In addition to that, it is not always possible to find close-range cameras in every place where the event occurs.

Due to the importance of the topic, many researchers tried to monitor behaviors of people using street, or indoor cameras which are also known as close-range cameras. However, most of the previous studies aimed to detect boundaries of large groups, and to extract information about them. The early studies in this field were developed from closed-circuit television images (Davies et al., 1995), (Regazzoni and Tesei, 1994), (Regazzoni and Tesei, 1996). Unfortunately, these cameras can only monitor a few square meters in indoor regions, and it is not possible to adapt those algorithms to street or airborne cameras since the human face and body contours will not appear as clearly as in close-range indoor camera images due to the resolution and scale differences. In order to be able to monitor bigger events researchers tried to develop algorithms which can work on outdoor camera images or video streams. Arandjelovic (Arandjelovic, Sep. 2008) developed a local interest point extraction based crowd detection method to classify single terrestrial images as crowd and non-crowd regions. They observed that dense crowds produce a high number of interest points. Therefore, they used density of SIFT features for classification. After generating crowd and non-crowd training sets, they used SVM based classification to detect crowds. They obtained scale invariant and good results in terrestrial images. Unfortunately, these images do not enable monitoring large events, and different crowd samples should be detected before hand to train the classifier. Ge and Collins (Ge and Collins, 2009) proposed a Bayesian marked point process to detect and count people in single images. They used football match images, and also street camera images for testing their algorithm. It requires clear detection of body boundaries, which is not possible in airborne images. In another study, Ge and Collins (Ge and Collins, 2010) used multiple close-range images which are taken at the same time from different viewing angles. They used three-dimensional heights of the objects to detect people on streets. Unfortunately, it is not always possible to obtain these multi-view close-range images for the street where an event occurs. Chao et al. (Lin et al., Nov. 2001) wanted to obtain quantitative measures about crowds using single images. They used Haar wavelet transform to detect head-like contours, then using SVM they classified detected contours as head or non-head regions. They provided quantitative measures about number of people in crowd and sizes of crowd. Although results are promising, this method requires clear detection of human head contours and a training of the classifier. Unfortunately, street cameras also have a limited coverage area to monitor large outdoor events. In addition to that, in most of the cases, it is not possible to obtain close-range street images or video streams in the place where an event occurs. Therefore, in order to behaviors of large groups of people in very big outdoor events, the best way is to use airborne images which began to give more information to researchers with the development of sensor technology. Since most of the previous approaches in this field needed clear detection of face or body features, curves, or boundaries to detect people and crowd boundaries which is not possible in airborne images, new approaches are needed to ex-

tract information from these images. Hinz et al. (Hinz, 2009) registered airborne image sequences to estimate density and motion of people in crowded regions. For this purpose, first a training background segment is selected manually to classify image as foreground and background pixels. They used the ratio of background pixels and foreground pixels in a neighborhood to plot density map. By observing change of the density map in the sequence, they estimated motion of people. Unfortunately, their approach did not provide quantitative measures about crowds. In a following study (Burkert et al., Sep. 2010), they used previous approach to detect individuals. Positions of detected people are linked with graphs. They used these graphs for understanding behaviors of people.

In order to bring a fully automatic solution to the problem, we propose a novel framework to detect people from remotely sensed images. One of the best solutions to monitor large mass events is to use airborne sensors which can provide images with approximately 0.3 m. spatial resolution. In previous studies (Sirmacek and Reinartz, 2011a) and (Sirmacek and Reinartz, 2011b), we used airborne images to monitor mass events. In the first study (Sirmacek and Reinartz, 2011a), we proposed a novel method to detect very dense crowd regions based on local feature extraction. Besides, detecting dense crowds, we have also estimated number of people and people densities in crowd regions. In following study (Sirmacek and Reinartz, 2011a), by applying a background control, individual persons are also detected in airborne images. Moreover, in a given airborne image sequence, detected people are tracked using Kalman filtering approach. Although airborne images are useful to monitor large events, unfortunately sometimes flying over mass event might not be allowed, or it might be an expensive solution. Therefore, detecting and monitoring crowds from satellite images can provide crucial information to control large mass events. As the sensor technology is being developed, new satellites can provide images with higher spatial resolutions. With those new satellite sensors, it became possible to notice human crowds, and even individual persons in satellite images. Therefore, herein we propose a novel approach to detect crowds automatically from very high resolution satellite images. Although resolutions of satellite images are still not enough to see each person with sharp contours, we can still notice a slight change of intensity and color components at the place where a person exists. Therefore, the proposed algorithm is based on local features which are extracted from intensity and color bands of the satellite image. In order to eliminate redundant local features which are generated by the other objects or texture on building rooftops, we apply a feature selection method which consists of three steps as; street classification approach, eliminating high objects on streets using shadow information, and using digital elevation model (DEM) of the region which is automatically generated using stereo satellite images to eliminate buildings. After applying feature selection, using selected local features as observations, we generate a probability density function (pdf). Obtained pdf helps us to detect crowded regions, and also some of the individual people automatically. We test our algorithm using Worldview-2 satellite images which are taken over Cairo and Munich cities. Our experimental results indicate the possible usage of the proposed approach in real-life mass events and to provide a rough estimation of the location and size of crowds from satellite data. Next, we introduce steps of the approach in detail.

2 LOCAL FEATURE EXTRACTION

In order to illustrate the algorithm steps, we pick $Munich_1$ image from our dataset. In Fig. 1.(a), we represent original $Munich_1$ panchromatic WorldView-2 satellite test image, and in Fig. 1.(b), we represent a subpart of this image in order to give information about real resolution. As can be seen here, satellite image resolutions do not enable to see each single person with sharp details. On the contrary, each person is represented with two or three mixed pixels, and sometimes additionally two or three mixed shadow pixels. All those pixels coming from a human appearance make a change of intensity components at the place where the person exists which can be detected with a suitable feature extraction method. Therefore, our crowd and people detection method depends on local features extracted from input image.

(a)

(b)

Figure 1: (a) $Munich_1$ test image from our Worldview-2 satellite image dataset, (b) Real resolution of a small region in $Munich_1$ test image.

For local feature extraction, we use features from accelerated segment test (FAST). FAST feature extraction method is especially developed for corner detection purposes by Rosten et al. (Rosten et al., Nov. 2010), however it also gives high responses on small regions which are significantly different than surrounding pixels. The method depends on wedge-model-style corner detection and machine learning techniques. For each feature candidate pixel, its 16 neighbors are checked. If there exist nine contiguous pixels passing a set of pixels, the candidate pixel is labeled as a feature location. In FAST method, these tests are done using machine learning techniques to speed up the operation. For detailed explanation of FAST feature extraction method please see (Rosten et al., Nov. 2010).

We assume (x_i, y_i) $i \in [1, 2, ..., K_i]$ as FAST local features which are extracted from input image. Here, K_i indicates the maximum number of features extracted from panchromatic band of the input image. We represent locations of detected local features for $Munich_1$ test image in Fig. 2.(b). As can be seen in this image, we have extracted local features on street at places where each individual person exits. Unfortunately, many redundant features are also detected generally on building rooftops, and corners. For detection of people and crowds, first of all local features coming from other objects should be eliminated. For this purpose, we apply a feature selection method that we represent in

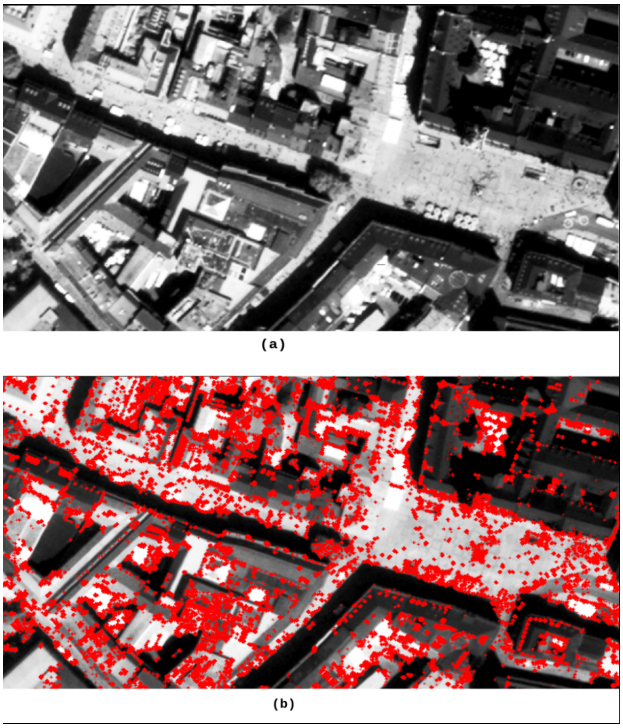


Figure 2: (a) Original $Munich_1$ test image, (b) FAST feature locations which are extracted from $Munich_1$ test image.

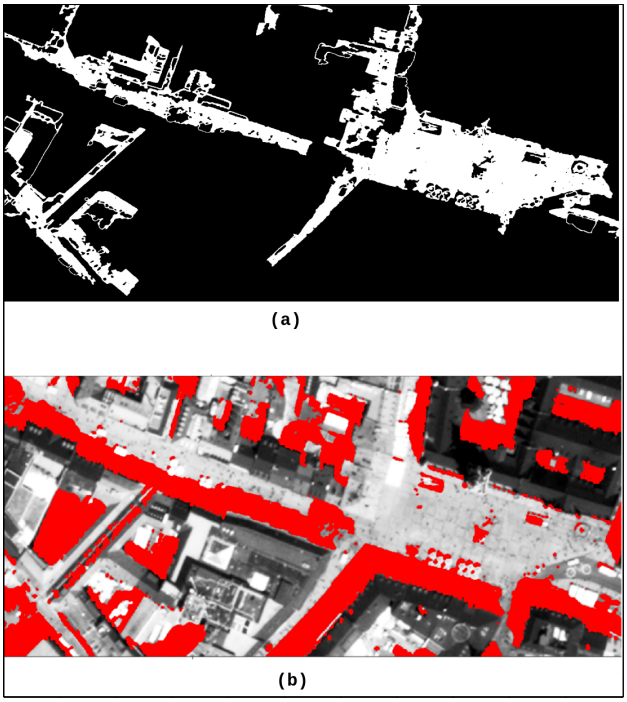


Figure 3: (a) Road-like pixels which are segmented from $Munich_1$ test image, (b) Automatically extracted shadow pixels from $Munich_1$ test image.

the next section in detail.

3 FEATURE SELECTION

For eliminating redundant features coming from building rooftop textures or corners of other objects in the scene, we use three masks as follows. The first mask ($B_1(x, y)$) is obtained by street segmentation using a training street patch which is selected by user. The second mask ($B_2(x, y)$), is generated using the shadow information, in order to remove high objects which appear on the detected street network. Finally, the third mask ($B_3(x, y)$) is obtained using height information obtained from DEM.

For street segmentation, we first choose a 20×20 pixel size training patch ($t(x, y)$) from input image. We benefit from normalized cross correlation to extract possible road segment. Normalized cross correlation between the training patch and the input image is computed using following equation.

$$\gamma(u,v) = \frac{\sum_{x,y}[g(x,y) - \overline{g}_{u,v}][t(x-u,y-v) - \overline{t}]}{\{\sum_{x,y}[g(x,y) - \overline{g}_{u,v}]^2 \sum_{x,y}[t(x-u,y-v) - \overline{t}]^2\}^{0.5}} \quad (1)$$

Here $\overline{t}$ represents the mean of intensity values in the template patch, and $\overline{g}_{u,v}$ represents the mean of the input image intensity values which are under the template image in correlation operation. At the normalized cross correlation result $\gamma(u, v)$, we obtain the road segment pixels as highlighted due to the high similarity to the training patch. By applying Otsu's automatic thresholding algorithm (Otsu, 2009) to the normalized cross correlation result, we obtain the road-like segments as in Fig. 3.(a). This binary image is assumed as the first mask ($B_1(x, y)$) which is going to be used for feature selection.

Although estimated street segment helps us for feature selection, still we cannot eliminate features coming from high objects on street such as street lamps, statues, small kiosks, etc. Unfortunately, those small objects also do not appear in DEM of the region, and they cannot be eliminated using height information coming from DEM. In order to eliminate features coming from these objects, in this step we try to detect them using shadow information. For shadow extraction, we use local image histograms. For each 100×100 pixel size window of the input image, the first local minimum in grayscale histogram is assumed as a threshold value to apply local thresholding to the image. After applying our automatic local thresholding method, we obtain a binary shadow map. In Fig. 3.(b), we represent detected shadow pixels on original image.

After detecting shadow pixels, we use the sun illumination angle to generate our high object mask. For labeling high objects, each shadow pixel should be shifted into opposite side of illumination direction. Assuming that (x_s, y_s) is an array of shadow pixel coordinates which are represented in Fig. 3.(b). New positions of shadow pixels $((\widehat{x}_s, \widehat{y}_s))$ are computed as $\overline{x}_s = x_s + l\sin(-\theta)$, and $\overline{y}_s = y_s + l\cos(-\theta)$. Here θ is the opposite direction of the illumination angle which is given by user, and l is the amount of shift in θ direction as pixel value. For better accuracy l should be chosen as the width of the shadow in illumination direction. However, in order to decrease computation time and complexity, we assume l equal to the length of the minor axis of an ellipse which fits shadow shape. After shifting shadow pixels, we generate our second mask $B_2(x, y)$ binary mask where $B_2(x, y) = 1$ for $((\widehat{x}_s, \widehat{y}_s))$. In Fig. 4, we illustrate shadow pixel shifting operation.

In order to obtain the last mask $B_3(x, y)$, we use DEM of the corresponding region which is generated from stereo Ikonos images using the DEM generation method of dAngelo et al. (dAngelo et al., 2009). We obtained $B_3(x, y)$ binary mask by applying local thresholding to DEM. We provide original DEM corresponding to $Munich_1$ image, and obtained binary mask in Fig. 5.(a), and (b) respectively. As can be seen, building rooftop regions are eliminated, however other low regions like park areas, parking

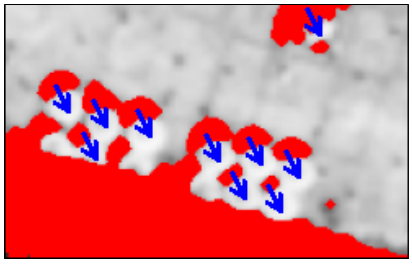


Figure 4: Illustration of shadow pixel shifting operation.

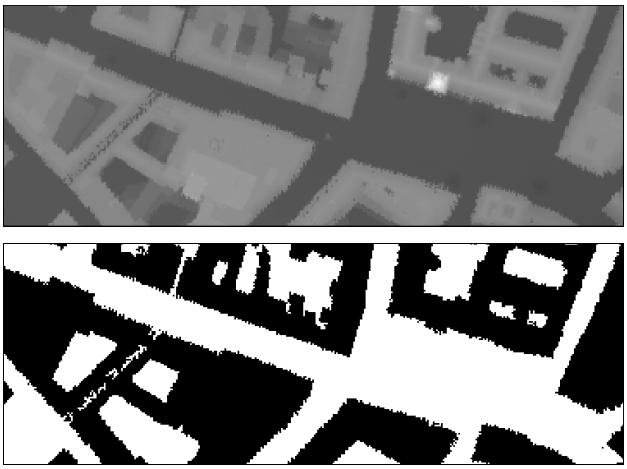


Figure 5: (a) Digital elevation model corresponding to $Munich_1$ test image which is generated using stereo WorldView-2 satellite images. (b) Low regions in $Munich_1$ image obtained by applying local thresholding to DEM.

lots with cars (or sea surface for some other test areas) cannot be eliminated with this mask. Therefore, we use information coming from three masks we generated. We assume our interest area as $S(x,y) = B_1(x,y) \wedge B_2(x,y)' \wedge B_3(x,y)$, where '$\wedge$' represents logical and operation for binary images.

We use detected $S(x,y)$ interest area for removing FAST features which are extracted from other objects. We eliminate a FAST feature which is at (x_i, y_i) coordinates, if $S(x_i, y_i) = 0$. Remaining FAST features behave as observations of the probability density function (pdf) of the people to be estimated. In the next step, we introduce an adaptive kernel density estimation method, to estimate corresponding pdf which will help us to detect dense people groups and also other people in sparse groups.

4 DETECTING INDIVIDUALS AND DENSE CROWDS

Since we have no pre-information about possible crowd locations in the image, we formulate the crowd detection method using a probabilistic framework. Assume that (x_i, y_i) is the ith FAST feature where $i \in [1, 2, ..., K_i]$. Each FAST feature indicates a local color change which might be a human to be detected. Therefore, we assume each FAST feature as an observation of a crowd pdf. For crowded regions, we assume that more local features should come together. Therefore knowing the pdf will lead to detection of crowds. For pdf estimation, we benefit from a kernel based density estimation method as Sirmacek and Unsalan represented for local feature based building detection (Sirmacek and Unsalan, 2010).

Silverman (Silverman, 1986) defined the kernel density estimator for a discrete and bivariate pdf as follows. The bivariate kernel function $[N(x,y)]$ should satisfy the conditions given below;

$$\sum_x \sum_y N(x,y) = 1 \tag{2}$$

$$N(x,y) \geq 0, \forall (x,y) \tag{3}$$

The pdf estimator with kernel $N(x,y)$ is defined by,

$$p(x,y) = \frac{1}{nh} \sum_{i=1}^{n} N(\frac{x - x_i}{h}, \frac{y - y_i}{h}) \tag{4}$$

where h is the width of window which is also called smoothing parameter. In this equation, (x_i, y_i) for $i = 1, 2, ..., n$ are observations from pdf that we want to estimate. We take $N(x,y)$ as a Gaussian symmetric pdf, which is used in most density estimation applications. Then, the estimated pdf is formed as below;

$$p(x,y) = \frac{1}{R} \sum_{i=1}^{K_i} \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma}) \tag{5}$$

where σ is the bandwidth of Gaussian kernel (also called smoothing parameter), and R is the normalizing constant to normalize $p_n(x,y)$ values between $[0, 1]$.

In kernel based density estimation the main problem is how to choose the bandwidth of Gaussian kernel for a given test image, since the estimated pdf directly depends on this value. For different resolution images, the pixel distance between two persons will change. That means, Gaussian kernels with different bandwidths will make these two persons connected to detect them as a group. Otherwise, there will be many separate peaks on pdf, however we will not be able to find large hills which indicate crowds. As a result, using a Gaussian kernel with fixed bandwidth will lead to poor estimates. Therefore, bandwidth of Gaussian kernel should be adapted for any given input image.

In probability theory, there are several methods to estimate the bandwidth of kernel functions for given observations. One well-known approach is using statistical classification. This method is based on computing the pdf using different bandwidth parameters and then comparing them. Unfortunately, in our field such a framework can be very time consuming for large input images. The other well-known approach is called balloon estimators. This method checks k-nearest neighborhoods of each observation point to understand the density in that area. If the density is high, bandwidth is reduced proportional to the detected density measure. This method is generally used for variable kernel density estimation, where a different kernel bandwidth is used for each observation point. However, in our study we need to compute one fixed kernel bandwidth to use at all observation points. To this end, we follow an approach which is slightly different from balloon estimators. First, we pick $K_i/2$ number of random observations (FAST feature locations) to reduce the computation time. For each observation location, we compute the distance to the nearest neighbor observation point. Then, the mean of all distances give us a number l. We assume that variance of Gaussian kernel (σ^2) should be equal or greater than l. In order to guarantee to intersect kernels of two close observations, we assume variance of Gaussian kernel as $5l$ in our study. Consequently, bandwidth of Gaussian kernel is estimated as $\sigma = \sqrt{5l}$. For a given sequence, that value is computed only one time over one

image. Then, the same σ value is used for all observations which are extracted from images of the same sequence. The introduced automatic kernel bandwidth estimation method, makes the algorithm robust to scale and resolution changes.

We use Otsu's automatic thresholding method on obtained pdf to detect regions having high probability values (Otsu, 2009). After thresholding our pdf function, in obtained binary image we eliminate regions with an area smaller than 1000 pixels since they cannot indicate large human crowds. The resulting binary image $B_c(x, y)$ holds dense crowd regions. Since our $Munich_1$ test image does not include very dense crowds, in Fig. 7 we illustrate an example dense crowd detection result on another Worldview-2 satellite test image which is taken over Cairo city when an outdoor event occurs.

After detecting dense crowds automatically, we focus on detecting individuals in sparse areas. Since they indicate local changes, we assume that detected local features can give information about individuals.

In most cases, shadows of people or small gaps between people also generate a feature. In order to decrease counting errors coming double counted people because of their shadows, we follow a different strategy to detect individuals. We use a binary mask $B_f(x, y)$ where (x_i, y_i) feature locations have value 1. Then, we dilate $B_f(x, y)$ using a disk shape structuring element with a radius of 2 to connect close feature locations. Finally, we apply connected component analysis to mask, and we assume mass center of each connected component as a detected person position. In this process, slight change of radius of structuring element does not make a significant change in true detected people number. However, an appreciable increase in radius can connect features coming from different persons which leads to underestimates.

5 EXPERIMENTS

To test the proposed algorithm, we use a Worldview-2 satellite image dataset which consists of four multitemporal panchromatic images taken over Munich city ($Munich_{1-4}$ images), and one panchromatic image taken over Cairo city ($Cairo_1$). Those panchromatic Worldview-2 satellite images have approximately half meter spatial resolution. We also test proposed algorithm on an airborne image (with 30 cm. spatial resolution) taken from the same region in over Munich city, in order to show robustness of the algorithm to resolution and sensor differences

In Fig. 6, we represent people detection results for $Munich_{1-4}$ images. For these four multitemporal images, true individual person detection performances are counted as $92,02\%$, $70,73\%$, $88,57\%$, and $89,19\%$ respectively. Besides, false alarm ratios are obtained as $14,49\%$, $40,34\%$, $24,29\%$, and $27,03\%$ respectively. In Fig. 7.(a), we present dense crowd detection and people detection results in Worldview-2 satellite image taken over Cairo city. Robust detection of dense crowd boundaries indicate usefulness of the proposed algorithm to monitor large mass events. Finally, in Fig. 7.(b), we represent people detection results on an airborne image which is taken in the same test area over Munich city. Obtained result proves robustness of the algorithm to scale and sensor differences of the input images.

6 CONCLUSION

In order to solve crowd detection and people detection, herein we introduced a novel approach to detect crowded areas automatically from very high resolution satellite images. Although resolutions of those images are not enough to see each person with sharp details, we can still notice a change of color components in the place where a person exists. Therefore, we developed an algorithm which is based on local feature extraction from input image. After eliminating local features coming from different objects or rooftop textures by applying a feature selection step, we generated a probability density function using Gaussian kernel functions with constant bandwidths. For deciding bandwidth of Gaussian kernel to be used, we used our adaptive bandwidth selection method. In this way, we obtained a robust algorithm which can cope with input images having different resolutions. By automatically thresholding obtained pdf function, dense crowds are robustly detected. After that, local features in sparse regions are analyzed to find other individuals. We have tested our algorithm on panchromatic Worldview-2 satellite image dataset, and also compared with an algorithm result obtained from an airborne image of the same test area. Our experimental results indicate possible usage of the algorithm in real-life events. We believe that, the proposed fully automatic algorithm will gain more importance in the near future with the increasing spatial resolutions of satellite sensors.

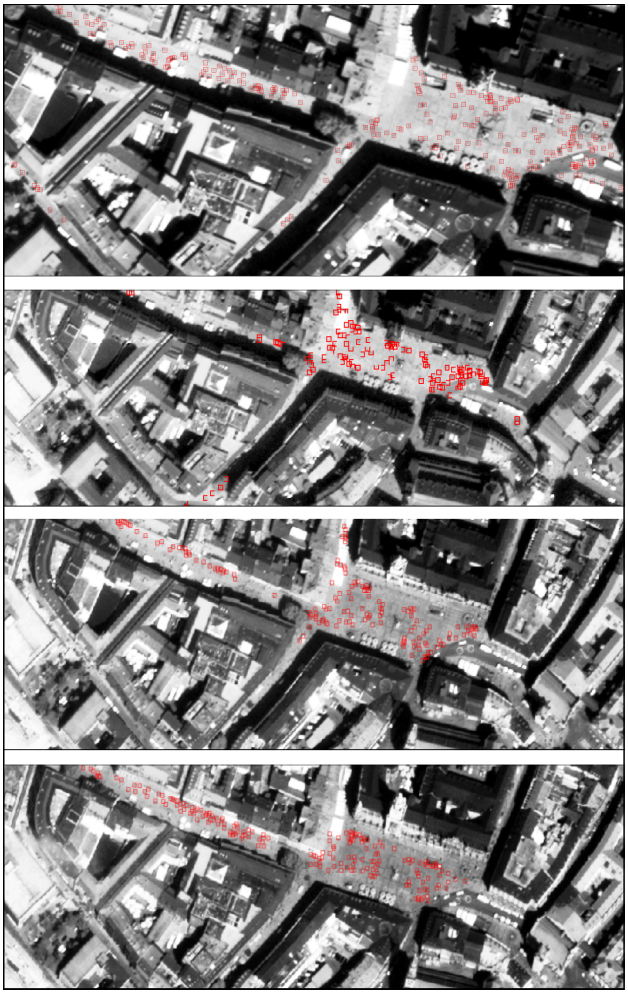
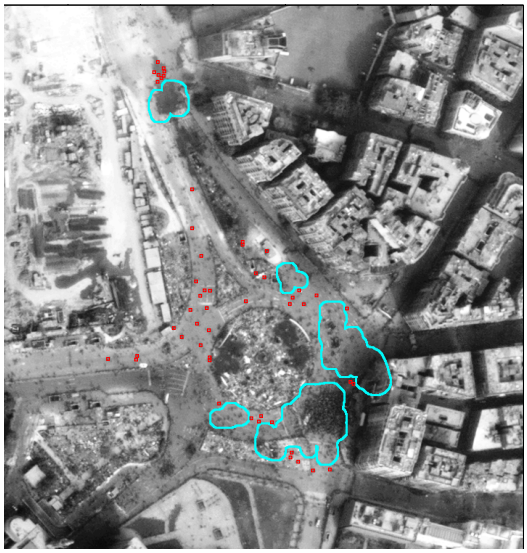


Figure 6: People detection results on $Munich_{1-4}$ Worldview-2 satellite images.

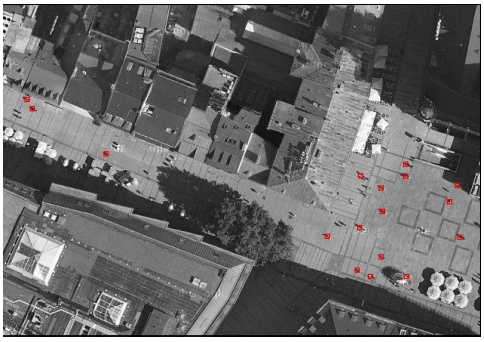
REFERENCES

Arandjelovic, O., Sep. 2008. Crowd detection from still images. British Machine Vision Conference (BMVC'08).

Burkert, F., Schmidt, F., Butenuth, M. and Hinz, S., Sep. 2010. People tracking and trajectory interpretation in aerial image sequences. International Archives of Photogrammetry, Remote Sensing and Spatial Infor-



(a)



(b)

Figure 7: (a) Dense crowd and people detection result on Worldview-2 satellite image taken over Cairo city, (b) People detection result on an airborne image which is taken at the same test area over Munich city.

mation Sciences (IAPRS), Commission III (Part A) XXXVIII, pp. 209–214.

dAngelo, P., Schwind, P., Krauss, T., Barner, F. and Reinartz, P., 2009. Automated dsm based georeferencing of cartosat-1 stereo scenes. In Proceedings of International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences.

Davies, A., Yin, J. and Velastin, S., 1995. Crowd monitoring using image processing. IEEE Electronic and Communications Engineering Journal 7 (1), pp. 37–47.

Ge, W. and Collins, R., 2009. Marked point process for crowd counting. IEEE Computer Vision and Pattern Recognition Conference (CVPR'09) pp. 2913–2920.

Ge, W. and Collins, R., 2010. Crowd detection with a multiview sampler. European Conference on Computer Vision (ECCV'10).

Hinz, S., 2009. Density and motion estimation of people in crowded environments based on aerial image sequences. ISPRS Hannover Workshop on High-Resolution Earth Imaging for Geospatial Information.

Lin, S., Chen, J. and Chao, H., Nov. 2001. Estimation of number of people in crowded scenes using perspective transformation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 31 (6), pp. 645–654.

Otsu, N., 2009. A threshold selection method from gray-level histograms. IEEE Transactions on System, Man, and Cybernetics 9 (1), pp. 62–66.

Regazzoni, C. and Tesei, A., 1994. Local density evaluation and tracking of multiple objects from complex image sequences. Proceedings of 20th International Conference on Industrial Electronics, Control and Instrumentation (IECON) 2, pp. 744–748.

Regazzoni, C. and Tesei, A., 1996. Distributed data fusion for real time crowding estimation. Signal Processing 53, pp. 47–63.

Rosten, E., Porter, R. and Drummond, T., Nov. 2010. Faster and better: A machine learning approach to corner detection. IEEE Transactions on Pattern Analysis and Machine Learning 32 (1), pp. 105–119.

Silverman, B., 1986. Density estimation for statistics and data analysis. 1st Edition.

Sirmacek, B. and Reinartz, P., 2011a. Automatic crowd analysis from airborne images. 5th International Conference on Recent Advances in Space Technologies RAST 2011, Istanbul, Turkey.

Sirmacek, B. and Reinartz, P., 2011b. Kalman filter based feature analysis for tracking people from airborne images. ISPRS Workshop High-Resolution Earth Imaging for Geospatial Information, Hannover, Germany.

Sirmacek, B. and Unsalan, C., 2010. A probabilistic framework to detect buildings in aerial and satellite images. IEEE Transactions on Geoscience and Remote Sensing.

Author Index