STRATEGIES FOR TEXTURING BUILDING MODELS WITH LOW RESOLUTION INFRARED IMAGE SEQUENCES

L. Hoegner [a], H. Kumke [a], A. Schwarz [b], L. Meng [a], U. Stilla [a]

[a] Institute of Photogrammetry and Cartography, Technische Universitaet Muenchen (TUM), 80290, Munich, Germany
{Ludwig.Hoegner; Holger.Kumke; Liqiu.Meng; Stilla}@bv.tum.de
[b] Research Institute for Optronic and Pattern Recognition (FGAN-FOM) , 76275 Ettlingen, Germany

KEY WORDS: Texturing, 3d models, image sequences, infrared, image analysis, texture extraction

ABSTRACT:

Focus of this paper is a strategies for texturing building models with partial images extracted from infrared image sequences. A given 3d building model is used for the projection of the images into the 3d object space. Caused by an insufficient camera position from the GPS the outer orientation of the virtual camera has to be corrected. Two different concepts are used depending on the geometric condition of the recorded scene and model. For scenes with many geometrical details visible in the corresponding 3d model view, the first concept is used. The polygons of the faces of the 3d model are projected into the IR image and matched to the edges for assigning the correct parts of the image to the corresponding faces of the model. These image parts are then projected onto the faces and stored as surface textures. For scenes with only few visible geometry in the projected 3d model, the second concept assumes a homography to detect surface planes of facades in two subsequent images and generates a camera path relative to those planes. Then, theses planes are matched to the 3d model to generate the final camera path. Generally, a single image does not show a complete building facade. The extracted textures of the images are combined to create complete textures for the model surfaces. For analyzing buildings these textures can be used for extraction of features or recognition of objects in a corrected image space.

1. INTRODUCTION

Thermal cameras are used for the detection of electro magnetic radiation invisible for normal cameras in the visible spectrum. Thermal cameras are able to resolve temperatures with an accuracy of 0,01 Kelvin for the recognition of objects with small differences in temperature and for identification of small details. High-quality infrared cameras are able to record image sequences with standard video frame rate (25 fps) or even higher. Thermal cameras are very expensive related to normal video cameras due to the special camera optics, the necessary cooling system to avoid thermal noise and the low production numbers. Today, thermal image data are used for many different applications. Satellite images are used for fire detection (Siegert, 2004), vegetation monitoring (Quattrochi, 1999) or the analysis of urban heat islands (Lo, 2003). Airborne IR-systems are applied for vehicle detection (Hinz, 2006, Stilla, 2002) or exploration of leakages in district heating systems (Koskeleinen, 1992). Ground cameras are recording the irradiation of building facades (Klingert, 2006), for the specification of its thermal behavior.

For urban environmental quality studies, satellite IR data were combined with 3d city models. In IR images are projected onto the terrain and the building surfaces involved in the energy exchange were matched with the IR images, including vertical walls and the horizontal surfaces of the buildings seen from the satellite. This textured model permit 3D visualization for a better understanding of the factors controlling urban environment (Nichol, 2005).

Typically, thermal inspection of single buildings is made directly and manually in photos. Bigger building parts are acquired by combining several images. So the results are marked in the 2d photos and stored as text without any geometric or 3d processing. This is a problem, when images from different cameras or views need to be combined and stored for further processing.

In this paper we describe automated strategies for texturing an building model or building complexes sited in inner city areas. Because of the low optical resolution of the IR cameras a small field of view of 20° is taken for the records. Due to this, it is necessary to record in an oblique view to see the complete height of the building. Several problems occur when trying to match the images and the model. Edges detected in the infrared image do not always have correspondences in the building model and vice versa This problem is even getting worse with a moving camera and inaccurate orientation parameters caused by the GPS system. Different strategies for matching of given 3d models and images are well known in computer vision. Single image processing is working with 3 or more correspondences between image and model. An overview over 3-point algorithms is given by Haralick (Haralick, 1994). Techniques for 4- and 5-point estimation are published by Quan (Quan, 1999)] and Triggs (Triggs, 1999). For 6 and more correspondence points the Direct Linear Transformation (DLT) can be used (Triggs, 1999). There are also iterative methods proposed (Haralick, 1989). Furthermore homographies can be used for plane detection (Hartley, 2000). When using image sequences, multiple images can be used for pose estimation. Longuet-Higgins (Longuet-Higgins, 1981)] uses the 8-point algorithm to reconstruct a scene from two different views.

2. CONCEPTS FOR TEXTURE EXTRACTION FROM IR IMAGE SEQUENCES

The proposed concepts are both based on the assumption that a 3d model of the recorded building is given with its vertices and

triangulated polygons. The coordinates should be given in a national coordinate system like Gauss-Krueger. For the projection of the 2d image pixel to the surfaces of the 3d model it is necessary to construct a virtual camera path with an outer orientation for every recorded image. The IR image sequence and the recorded GPS-based camera orientation are synchronized via a time stamp. Depending on the number of 3d model details, especially edges and vertices that can be seen different strategies have to be preferred. For images with complete 3d model surfaces within the field of vision, like airborne sequences, a direct matching can be performed using edge detection and spatial intersection to correct the initial camera position. If not enough edges and vertices of the 3d model are within the image, the subsequent IR images are used to generate planes from the images via homographies to estimate a camera position relative to those estimated planes. The local coordinate system generated from the homographies is translated to global coordinates by matching the estimated planes from the image pairs and the surfaces of the 3d model in world coordinates.

2.1 Camera correction and matching of images and 3d model

The image sequence consists of two types of images, key images with a time stamp corresponding to a recorded camera orientation, and images without a corresponding camera orientation. For the key images of a sequence, only the recorded exterior orientation of the IR camera is given. This position is not exact enough for the extraction of the textures of the building surfaces, so it must be corrected. Edge detection is performed in the IR image. Here, a simple Canny edge detection is used. Then, for the polygon edges of the model the corresponding edges in the IR image are searched. This is only possible, if the virtual camera sees enough model edges to uniquely match the scene. For this, at least three visible vertices of the 3d model are necessary for spatial intersection. This depends on the level of detail of the given 3d building model and the camera view. If there are not enough edges and vertices of the 3d model corresponding to visible edges in the IR image, the camera position cannot be corrected. At least three vertices are necessary for this correction. These corrected positions at the key images are used for linear interpolation of the virtual camera for the images between the key images.

When the number of visible vertices of the 3d model is insufficient, it is necessary to follow a different strategy. Because the points of facades are approximately building a plane, it is possible to calculate a homography matrix for every two subsequent images with

$$p_2 = K_2 \cdot H \cdot K_1^{-1} \cdot p_1 \qquad (1)$$

where $H = R + (t \cdot (n^T / d))$

R = rotation matrix
t = translation vector
n = normal vector of the plane
d = distance to the plane
K1 and K2 = calibration matrices for the cameras

For estimation of H for the first image of the pair, corners with large eigenvalues are detected. First, the minimal eigenvalue for every pixel is calculated. Then non-maxima suppression is performed in 3x3 neighborhoods. Points beneath a threshold and points to close to another point of interest are rejected. For every selected point in the first image, the second image is scanned for a corresponding point with significant eigenvalue. Pairs of corresponding points of interest are stored. The homography matrix is calculated from pairs of points selected from all stored points using RANSAC. If there is more than one façade within the image, the image can be subdivided at the detected edges. In this case points of interest are stored separately for different supposed planes.
After the estimation of the homographies of all planes of the image pair, the relative rotation and translation of the camera from the first to the second image can be calculated. This is done for all subsequent images of the image sequence. In the end, there are estimated planes for the facades and a corresponding camera path. The coordinates of the planes and the camera path are relative to the first camera position, which is set to the origin looking along the positive z axis. To transfer the camera and the planes to the correct 3d coordinates of the building model the plane equations generated from the homography are compared to plane equations calculated from the polygon surfaces of the 3d model. Corresponding pairs of planes in the image and the 3d model are searched and verified by comparing the recorded camera positions to those, created in the homographies.

2.2 Extraction of partial textures for visible surfaces

After the correction of the camera orientation the image points have to be projected onto the model surfaces. First, it is necessary to determine the visibility of model surfaces in the image. This is done by ray casting (Foley, 1995). The triangles of the polygons are clipped at the near plane and far plane of the virtual camera. Triangles that are completely outside the left, right, top and bottom plane are also removed. For all remaining triangles, their plane equation is calculated from their vertices. For every pixel of the IR image, a ray is sent trough the image plane to detect its intersection points with all triangle planes. Starting with the intersected plane with the lowest depth value, the intersection point is checked to be inside the edges of the surface. This is done by using barycentric coordinates. Barycentric coordinates define the barycenter of a triangle as point (0,0,0). The three vertices of the triangle have the coordinates (1,0,0), (0,1,0) and (0,0,1). Every point is then described by

$$(a_1 + a_2 + a_3) \cdot p = a_1 x_1 + a_2 x_2 + a_3 x_3 \qquad (2)$$

where p = point
a1, a2, a3 = barycentric coordinates of p
x1, x2, x3 = vertices of the triangle

If a1, a2 and a3 are positive and a1 + a2 + a3 = 1, point p is inside the triangle. The visible plane for the pixel is the intersected plane with the lowest depth value, where the intersection point is within. Then, the pixel is transferred from 2d image coordinates to 3d model coordinates.

Next, the position of the pixel on the surface triangle has to be calculated. For the vertices that define the triangle, their texture coordinates are known. Generating two plane equations for the x and y value of the texture coordinate, for every point

projected onto the triangle its texture coordinates can be determined. Now, the image pixels are projected from 3d model space to 2d texture space and transferred from image pixels to texture elements (texels).

By using the 2d texture coordinates and the 3d model coordinates, the spatial resolution of a surface texture extracted from one IR image can be calculated. The mean distance between one texel and its neighbors in 3d model space is divided by the mean distance in 2d texture space. For every surface triangle every texel extracted from one image is stored with its value, texture coordinates, spatial resolution and the ID of the image it was taken from. These texels of one image can be used to generate a partial surface texture.

2.3 Combination of the partial textures of one surface

For composing a complete texture for a building façade, it is necessary to combine several partial textures. For every pixel of the final texture its value is interpolated from the stored texels. First, texture coordinates for the pixel are calculated. Next, the partial textures stored as texels, like described above, are searched for the best texels to calculate the pixel's value. Best means texels from images of the same sequence, high spatial resolution and with a small distance to the texture coordinate of the pixel. From those texels the ones with the highest spatial resolution are taken to interpolate the value of the pixel of the final surface texture. Because the appearance of IR signatures is changing over time and depends for example on the weather conditions, the combination of textures to create a new combined texture is very difficult. That is why normally only textures of one record session should be used to create new textures.

The textures, are assigned to building surfaces and stored together with the 3d model in a hierarchical object oriented database for further processing like feature extraction.

3. DATA ACQUISITION

As the wavelength of infrared light is much longer than the visible spectrum, other optics and sensors are necessary to record infrared radiance. Current IR cameras cannot reach the optical resolution of video cameras or even digital cameras. The cameras that were used for the acquisition of the test sequences offer an optical resolution auf 320x240 (FLIR SC3000) and 320x256 (MerlinMID) pixel with a field of view (FOV) of only 20°. The SC3000 is recording in the thermal infrared (8 - 12 µm), whereas the MerlinMID is a midwave infrared camera (3 - 5 µm). On the top of a van, both cameras were mounted on a rotatable and shiftable platform. Images in the mid-wave infrared are directly affected by the sunlight as in addition to the surface radiation caused by the building's temperature the radiation of the sun is reflected. In long-wave infrared the sun's influence appears only indirect, as the sun is not sending in the long wave spectrum, but of course is affecting the surface temperature of the building.

To inspect the temperature loss of a building, the environment should be cold, where as the heating of the building should be running to observe a relevant temperature flux. Optimal time is after sunset and before dawn in spring. To minimize occlusion caused by vegetation a date before foliation is to be preferred.
Caused by the small field of view and the low optical resolution it was necessary to record the scene in oblique view to be able to record the complete facades of the building from the floor to the roof and an acceptable texture resolution. That is why the optical resolution of the images is not constant. Image parts that show the complete building have a lower resolution than image parts that show only the first floor of the building. When several images are combined, the optical texture quality of the final surface texture is high in the first floor and decreases to the top auf the façade. Due to this, the image sequences were recorded with a frequency of 50 frames per second. To minimize holes in the textures due to occlusion caused by the oblique view, every façade was recorded with a view forward looking and a view backward looking. The viewing angle related to the along track axis of the van was constant. The position of the cameras was recorded with GPS and, for quality measurements from tachymeter measurements from ground control points.

4. PROCESSING OF TEST DATA

As mentioned in chapter II, different strategies for the matching are necessary dependent on the objects, which can be seen by the camera. Using data recorded as mentioned in chapter III both matching strategies and the texture extraction and combination are examined.

4.1 Camera correction and matching of images and 3d model

The direct matching method with point correspondences between an image and the 3d model projection can be used when at least 3 vertices of the building model are visible. So, it is usable for airborne records, as those images show a lot of polygons of building models. On the other hand, when the camera is close to a building façade, often only one surface can be seen and even this surface is not completely projected into the image plane with its edges and vertices. In this case the edge detection and matching provide no correct camera position. This error can be seen in Figure 1 and Figure 2. In Figure 1 the image is matched with the correct edges but wrong scale. Because the corners of the façade are not within the image, it is not possible to determine the height of the façade as well as the length of the ground line.

Figure 1. At first glance the polygons fit well, but the scale cannot be estimated because of missing vertices to correct the camera position

In Figure 2 the next IR image is loaded with the estimated 3d model view as overlay. Due to the wrong camera position in Figure 1 the path estimation for Figure 2 leads to a wrong matching.

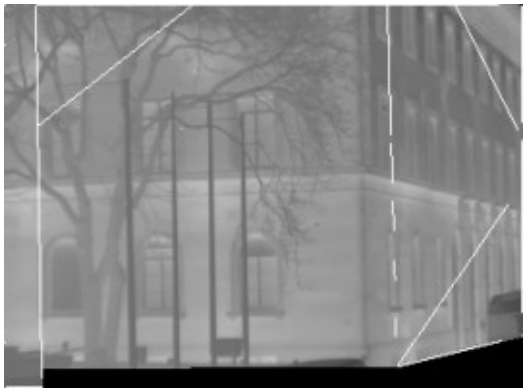

Figure 2. The subsequent image of the image used in Figure 1

Figure 3 shows the same image used in Figure 1 overlayed with arrows indicating the movement of points of interest in between the image and the subsequent image (Figure 2). Out of these points RANSAC chooses pairs of points for the plane estimation. If a situation occurs like in Figure 3, where there is more than one façade within the image, the image can be subdivided at the detected edges or by RANSAC, when dividing point pairs to different planes when there estimated movement is not the same.

Figure 3. IR image with selected points of interest, that have correspondences in the following image. Arrows show the moving direction of the point

4.2 Texture Generation

there is a corrected camera orientation that is used for the virtual camera, that project the image pixels into the 3d model space. As mentioned in chapter II section B, raycasting is uned to determine the visible surface for every pixel and then calculate its texture coordinates. From these texels of one image a partial surface texture for every visible surface is generated as seen in Figure 4. Areas of the surface, which could not be seen in one image, are left in black. As this texture is extracted from an oblique view, the spatial resolution is not constant. It is getting lower from the left to the right.

Figure 4. Partial texture of one façade extracted from the IR image in Figure 1 and Figure 2

For composing a complete texture for a building façade, for every pixel of the final texture its value is interpolated from the stored texels. This procedure metioned in chapter II section C is very time and memory consuming. A possible way to reduce complexity is to use knowledge about the records. When the camera is forward looking, to the left then the left border of every image of the sequence will have the highest resolution. Figure 5a gives an idea about the resolution within a partial texture given by a single image of the sequence. The distribution of the resolution is coded by different grey values. All partial textures are sorted with the ascending ID of their origin images. Every texture of the following images is moving rightwards. Figure 5b shows the resulting distribution of the resolution given by a combination of several textures extracted from one image sequence. Every following texture overwrites a part of the previous texture with a higher resolution. Because higher resolution can only be obtained from the first floor of the building, the upper building parts remain in the lower resolution of the previous image.

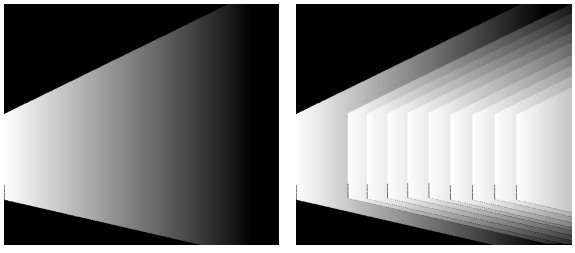

Figure 5. Resolution images of the surface shown in Figure 4 White means high resolution, black is low resolution, a) Resolution image of a single partial texture, b) Resolution image of a combination of several partial textures.

Generating the final texture in this way leads to a texture with the highest resolution for every texel that is extractable from all partial textures. Figure 6 shows a continuous distribution of the resolution of the complete texture of the façade assuming a high image rate and a low velocity.

Figure 6. Resolution image of the final texture

The white region in the middle corresponds to the first floor of the building, which lies in the view axis of the camera and can be seen at the left high resolution edge of all partial textures. Due to the oblique view the resolution decreases downwards and upwards. These texture parts cannot be seen at the left edge of the images, but more to the right, where the resolution decreases (seen Figure 5a). This effect depends on the angles of the camera view towards the camera path and towards the facades. The bigger the angle between the viewing direction of the camera and the surface normal of the façade, the bigger gets this effect. For the scenario described in this paper, texture generation can be done separately for the sequence in oblique forward view and for the sequence in oblique backward view. After that, both resulting textures can be combined. Figure 7 shows a surface texture calculated from three selected IR images, tow in forward view and one in backward view.

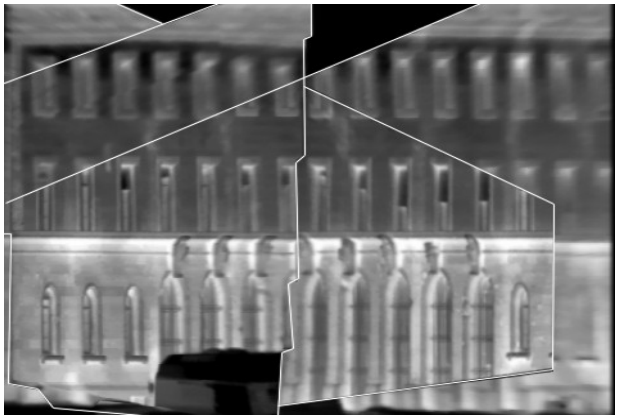

Figure 7. Surface texture generated from 3 partial textures. The edges between the taken textures are highlighted

5. DISCUSSION

Infrared light has special characteristics that lead to many problems. The physical behavior of the IR spectrum causes camera systems with lower resolution than normal video or photo cameras. The appearance of a façade in the depends on outer conditions like weather, daytime, sun, temperature and inner conditions like building structure or heating. Many things to be seen in one infrared image, can not be seen in another because the conditions have changed and many façade structures that can be seen in visible can not be seen in infrared and vise versa. In addition, the small viewing angle and resolution of the camera does not allow the record of a complete building in dense urban areas with narrow streets in one image. As mentioned in chapter 4, two different strategies were chosen to get surface related textures from the IR images. The first and direct way to project the image into the 3d model space is only possible, when the position of the IR camera can be determined correctly and accurately. As GPS is not accurate enough, it is necessary to correct the camera position for the projection. But this is only possible, if there are at least 3 corresponding points in the IR image and the 3d model image projection. That is the case, when the model has a high level of detail or the camera sees several facades with their edges and corners for example from airborne sequences.

Another problem of this matching is the difference between edges detected in the IR image and edges in the 3d model. The roofs in 3d models normally are not modeled with a roof overhang. But IR images from the street see the inner and the outer edge of the roof overhang. Due to this, the overhang that can be seen in the IR image is projected onto the façade which causes a distortion in the façade texture.

The second matching strategy with homography avoids those problems. Edges do not need to be present with corners as they are only used to subdivide the image to detect several planes. But this homography is only successful for images with few surfaces to detect. Therefore, the facades must contain enough details to extract points of interest for the homography. Point pairs, that are not within the plane, like trees, falsify the homography and are rejected by RANSAC. The oblique view caused by the viewing angle and resolution of the cameras leads to self occlusion of buildings. This is partially solved by recording two different views per façade, one in oblique forward view and one in oblique backward view. But, there are still remaining parts of buildings that are not visible.
A result of the texture combination, where two views in oblique forward were integrated and afterwards combined with one oblique backward view, is given in Figure 8.

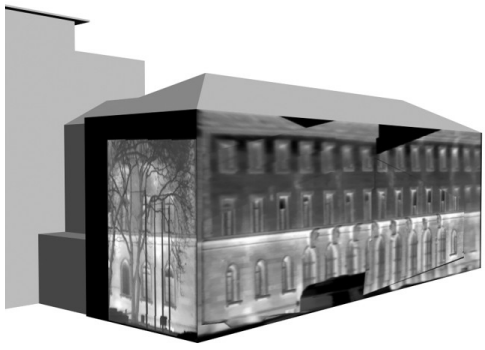

Figure 8. Model of one building of the TUM building complex. The black triangles on the right façade are not visible in the three used images

ACKNOWLEDGMENT

Our work is part of the DFG (German Research Society) research project “Enrichment and Multi-purpose Visualization of Building Models with Emphasis on Thermal Infrared Data“

(STI 545/1-2) as part of the bundle project “Interoperation of 3D Urban Geoinformation (3DUGI)”. The authors thank Dr. Clement, Dr. Schwarz and Mr. Kremer of FGAN-FOM, Department Optronic Systems Analyses, Ettlingen, for their assistance during the recording campaign.

REFERENCES

Foley, J..D., 1995. *Computer Graphics: Principles and Practice*, Addison-Wesley, ISBN 02-018-4840-6

Haralick, R.M., Joo, H., Lee, C.N., Zhuang, X., Vaidya, V.G., Kim, M.B., 1989. Pose estimation from correspondence point data, *SMC*, vol.19(6), pp. 1426-1446

Haralick, R.M., Lee, C.N., Ottenberg, K., Nolle, M, 1994. Review and analysis of solutions of the 3-point perspective pose estimation problem, *IJCV*, vol.13(3), pp. 331-356

Hartley, R.L., Zisserman, A, 2000. *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521623049

Hinz, S., Stilla, U, 2006. Car detection in aerial thermal images by local and global evidence accumulation, *Pattern Recognition Letter*, vol. 27, pp. 308-315

Klingert, M., 2006. The usage of image processing methods for interpretation of thermography data “,17th International Conference on the Applications of Computer Science and Mathematics in Architecture and Civil Engineering, Weimar, Germany, 12-14 July 2006

Koskeleinen, L., 1992. Predictive maintenance of district heating networks by infrared measurement, *Proc. SPIE*, vol. 1682, pp. 89–96

Lo, C.P., Quattrochi, D.A., 2003. Land-Use and Land-Cover Change, Urban Heat Island Phenomenon, and Health Implications: A Remote Sensing Approach, *Photogrammetric Engineering & Remote Sensing*, vol. 69(9), pp. 1053–1063

Longuet-Higgins, H.C., 1981. A computer algorithm for reconstruction a scene from two projections”, *Nature*, vol.239, pp. 133-135

Nichol, J., Wong, M.S., 2005. Modeling urban environmental quality in a tropical city, *Landscape and urban Planning,* vol.73, pp.49-58

Quan, L., Lan, L.D., 1999. Linear n-point camera pose determination, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.21(8), pp. 774-780

Quattrochi, D.A., Luvall, J.C., 1999. Thermal infrared remote sensing for analysis of landscape ecological processes: methods and application, *Landscape Ecology* , vol. 14, pp. 577–598

Siegert, F., Zhukov, B., Oertel, D., Limin, S., Page, S.E., Rielay, O., 2004. Peat fires detected by the BIRD satellite, *International Journal of Remote Sensing*, vol. 25(16), pp. 3221-3230

Stilla, U., Michaelsen, E., 2002. Estimating vehicle activity using thermal image sequences and maps, Symposium on geospatial theory, processing and applications. *International Archives of Photogrammetry and Remote Sensing*, vol.34,Part 4

Triggs, B., 1999. Camera pose and calibration from 4 or 5 known 3d points, *Proc. International Conference on Computer Vision (ICCV’99)*