MODEL–BASED 3D SCENE ANALYSIS FROM STEREOSCOPIC IMAGE SEQUENCES

Reinhard Koch
Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung
Universität Hannover, Appelstrasse 9A, 3000 Hannover 1, Germany
ISPRS Commission V

An approach for the modeling of complex 3D scenes like outdoor street views from a sequence of stereoscopic image pairs is presented. Starting with conventional stereoscopic correspondence analysis a 3D model scene with 3D surface geometry is generated. Not only the scene geometry but also surface texture is stored within the model. The 3D model permits to detect and correct geometric errors by comparison of synthesized images with real input images through analysis by synthesis techniques. 3D camera motion can be estimated directly from the image sequence to track camera motion and to fuse measurements from different viewpoints throughout the sequence into a common 3D model scene. From the textured 3D model realistic looking image sequences from arbitrary view points can be synthesized using computer graphics methods.

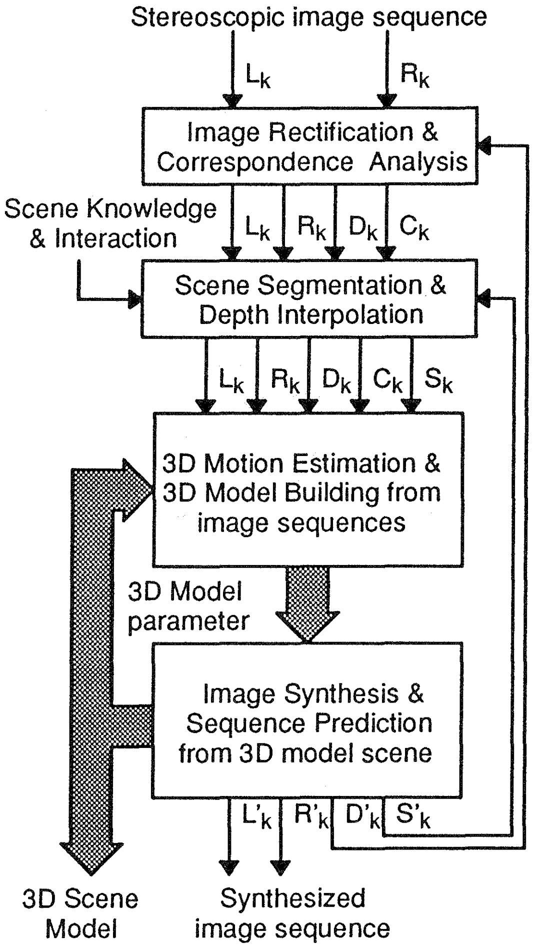
Key Words: 3D Scene Analysis, Stereoscopic Image Sequence Analysis, Robot Vision, Scene Reconstruction, Close Range Photogrammetry.

INTRODUCTION

Conventional stereoscopic image analysis tries to reconstruct a 3D scene from pairs of camera images through triangulation of corresponding 2D image points while the relative orientation of the cameras to each other is known. By triangulation a depth map can be constructed where the distance of each corresponding 3D scene point to the camera focal point is measured. This approach is sufficient for simple scene geometry without occluded areas but will fail when analyzing complex scenes like street views.

The reconstruction of complex 3D scenes requires that a series of problems need to be solved, especially the problems of 2D correspondence and 3D registration. To be able to triangulate the scene from a particular pair of images a calibration of the projection geometry and the relative orientation of the cameras to each other is needed. When using a binocular camera setup, the calibration can be performed once before the measurement and will remain constant throughout the measurement phase. Measurement of scene geometry relative to such a binocular camera system can then be obtained through triangulation of corresponding image points. During correspondence analysis one tries to uniquely identify the projections of a scene element onto the camera targets. In a complex natural scene parts of the scene may be occluded to the camera system so that the camera has to be moved throughout the scene. The scene is then analyzed from a sequence of image pairs. To register all measurements into a common scene coordinate system the 3D motion of the camera system must be tracked and measurements from multiple view points must be integrated to build a 3D model of the scene [Aloimonos, 1989].

The presented approach addresses the problems stated above for building a 3D model of a complex scene from a sequence of stereoscopic image pairs. Fig. 1 displays the structure of a 3D scene analysis system that automatically extracts 3D shape, motion, and surface texture of a 3D scene viewed by a stereoscopic camera. Input to the system is a stereoscopic


Stereoscopic image sequence
L_k
R_k
Image Rectification & Correspondence Analysis
Scene Knowledge & Interaction
L_k R_k D_k C_k
Scene Segmentation & Depth Interpolation
L_k R_k D_k C_k S_k
3D Motion Estimation & 3D Model Building from image sequences
3D Model parameter
Image Synthesis & Sequence Prediction from 3D model scene
L'_k R'_k D'_k S'_k
3D Scene Model
Synthesized image sequence

L_k : left image
R_k : right image
D_k : disparity map
C_k : confidence map
S_k : segmantation map
$_k$: Frame k of sequence

Fig. 1: Structure of 3D scene analysis.

image sequence and the known camera parameters of the stereoscopic camera pair. In a first step the images are rectified and corresponding points are searched in each image pair of the sequence. For each image pixel an estimate of image disparity is calculated and stored in a disparity map D_k together with a confidence measure that describes the quality of the estimate in C_k. The local disparity measurements are merged to physical objects during scene segmentation and the physical object boundaries are recorded in a segmentation map S_k. Prior knowledge of the observed scene as well as human interaction that guides the segmentation process can be included to improve the modeling quality. All measurements of one object are interpolated to smooth object surfaces and to fill gaps in the depth map. in the scene segmentation and interpolation stage. All information obtained so far from image pair analysis are fused in a 3D scene model. The disparity map is converted into a depth map and a 3D surface description is derived from the depth measurements. The surface geometry is represented as a triangular surface mesh spanned by control points in space. These control points can be shifted to adapt the surface geometry throughout the sequence. Not only the scene geometry but also the scene surface texture is stored within the model. It is therefore possible to synthesize realistic looking image sequences (L'_k, R'_k) from the model scene using 3D computer graphics methods [Koch, 1990]. A 3D motion estimation algorithm is included that calculates the motion of the camera and object motion throughout the scene and allows to fuse measurements from multiple view points. From the model scene predictions of the measurements (D'_k, S'_k) can be calculated together with the synthesized sequence (L'_k, R'_k) and used in a feedback loop to further enhance the reliability of the measurements. This feedback loop improves the 3D scene analysis based on comparison of the synthesized 2D sequence with the real image sequence based on the analysis by synthesis principle.

STEREOSCOPIC IMAGE PAIR ANALYSIS

The analysis of a stereoscopic image pair is split into correspondence analysis and scene segmentation. The correspondence analysis tries to locally estimate image plane correspondences while during scene segmentation image areas that belong to physically connected regions are identified through similarity measures and merged to scene objects. In a preprocessing step the image pair is rectified to give an image pair where the camera axes are parallel and the cameras are displaced is in horizontal image plane coordinates only. This image rectification greatly simplifies correspondence analysis and the search space is reduced to parallel horizontal epipolar lines **E**.

Correspondence analysis

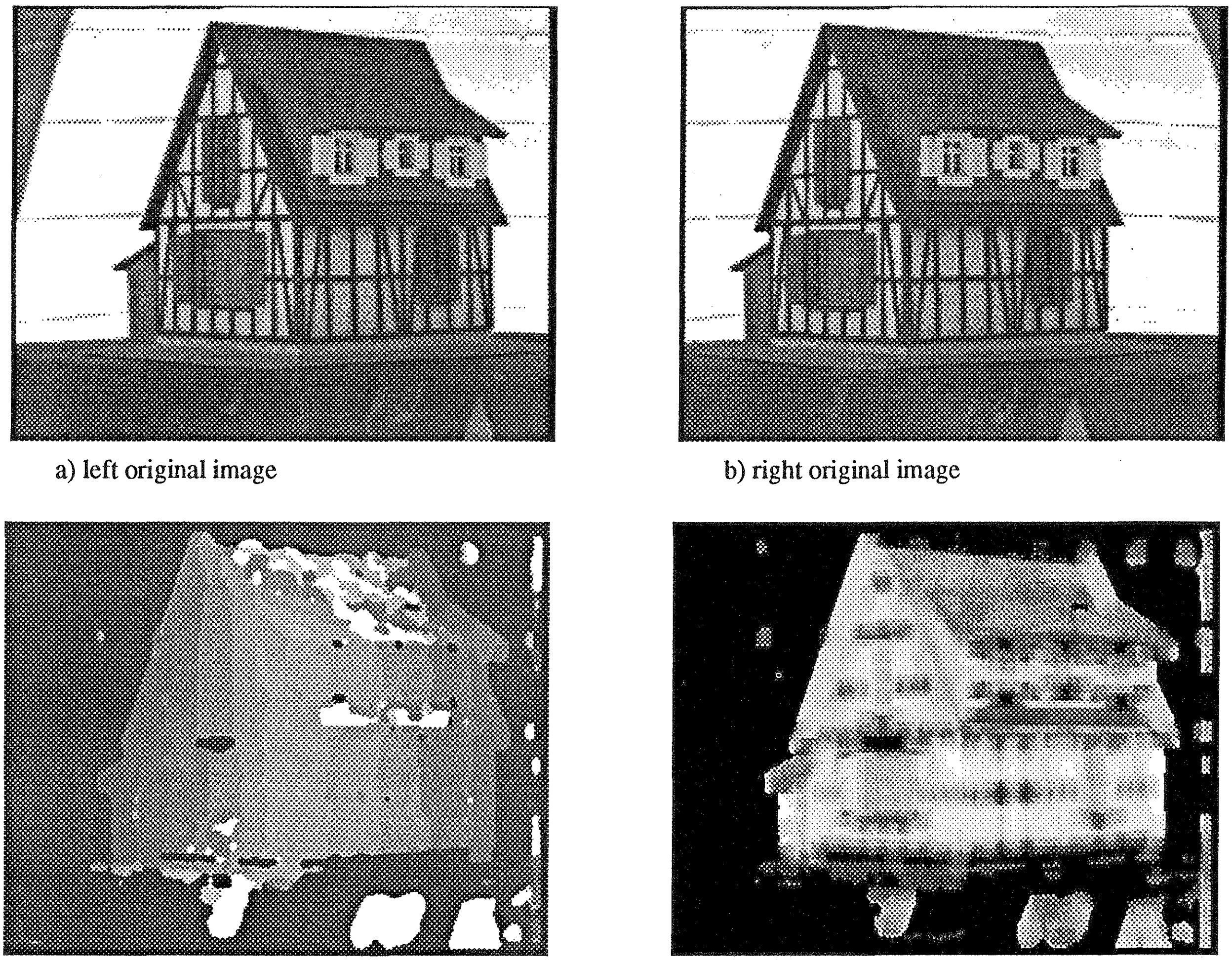
The correspondence analysis is split in three parts. First a candidate for a corresponding point must be identified in one image, then the corresponding candidate in the other image is searched along the epipolar lines and third the most probable candidate match between both images is selected based on a quality criteria. This search is repeated for each candidate, that is for each pixel. To select candidates the image grey level gradient **G** is evaluated. The image gradient is a vector field pointing into the direction of changing image texture like grey level edges. Only areas exceeding a minimum image gradient value $|\mathbf{G}| > G_{min}$ can be candidates for correspondence. The quality of the candidate can be estimated when comparing the gradient direction with the search direction. Edges perpendicular to the search direction can be located best while edges parallel to the search direction cannot be located at all. This quality measure C_1 can be calculated in Eq. (1). Candidates with $C_1 = 0$ can not be estimated there candidates with $C_1 = 1$ have highest confidence in estimation.

$$C_1 = \left\{ \begin{array}{ll} 0 & \text{for} \, |\mathbf{G}| < G_{min} \\ \dfrac{\mathbf{G} \cdot \mathbf{E}}{|\mathbf{G}|} & \text{else} \end{array} \right\} \qquad (1)$$

The estimation of C_1 is carried out for each image pixel. Each pixel with a gradient quality measure of $C_1 > 0$ will be selected as candidate. For each candidate a small measurement window (typically 11*11 pixel) around the candidate position in one grey level image is chosen and the corresponding grey level distribution is searched for in the other image. The search space is reduced to a one–dimensional search along the epipolar line between minimum and maximum disparity values derived from the known minimum and maximum scene distance. To select the most probable corresponding candidate along the search line, the normalized cross correlation (NCC) is calculated between the candidates. The most probable candidate pair is the pair with maximum cross correlation. The disparity value obtained for this candidate pair is recorded in a disparity map. The NCC is additionally used to define the correspondence quality. Selected corresponding pairs with low NCC are corresponding points with low confidence. Therefore a second quality measure C_2 in Eq. (2) can be defined that reflects the correspondence measurement confidence. Experiments have shown that candidates below a minimum threshhold NCC_{min} (NCC_{min} being approximately 0.7) are most often false matches that should be discarded. The confidence quality is therefore defined to be zero below NCC_{min} and NCC elsewhere.

$$C_2 = \left\{ \begin{array}{ll} 0 & \text{for} \, NCC < NCC_{min} \\ NCC & \text{else} \end{array} \right\} \qquad (2)$$

Both quality measures can be merged to one measure $C_c = C_1 \cdot C_2$ that contains the combined quality measure for each candidate. From the correspondence analysis two candidate maps are created: a disparity map contains the most probable disparity value for each candidate and a confidence map con-



a) left original image

b) right original image

c) disparity map (dark = far from camera, light = near to camera)

d) confidence map (dark = low confidence, light = high confidence)

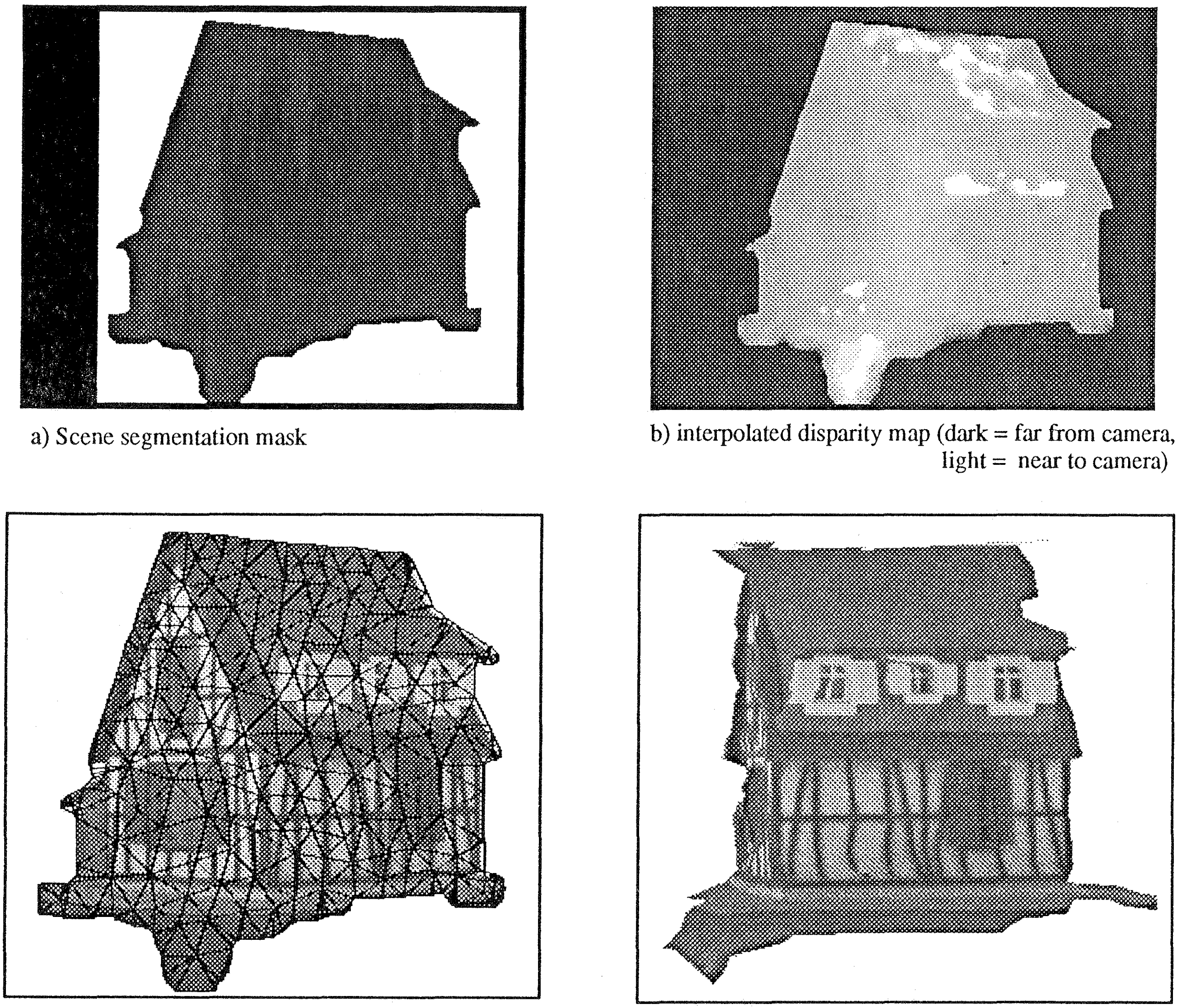
Fig. 2: Correspondence analysis of image pair "house".

tains the combined quality measure C_c for each candidate. Fig. 2 demonstrates the correspondence analysis for an image pair of the sequence "house". The image sequence "house" consists of a series of 90 views of the house where the house is rotated 4 degrees around the vertical axis in each view. The cameras are displaced 15 mm in horizontal direction with parallel optical axes and the house is placed approximately 400 mm from the camera origin. Fig. 2a and b show the left and right input image, Fig. 2c the disparity map and Fig. 2d the corresponding confidence map. Black regions are regions where no disparity could be measured. The measured disparity values are between 30 and 50 pixel. It can be seen that some regions in the area of the roof have false disparity values. This areas correspond to regions with low confidence because no surface structure is available to uniquely select a candidate.

Scene segmentation

The correspondence analysis yields a disparity map based on local depth measurement only. These measurements are uncertain and must be merged to regions that describe physical object surfaces. Based on similarity measures of scene depth the segmentation divides the viewed scene into contiguous surfaces and merges all disparity measurements of one object surface. The object boundaries are corrected from the grey level image with a contour approximation by assuming that physical object boundaries most often create grey level edges in the image.

Once the depth map is segmented into object regions all measurements of one region are interpolated by a thin plate surface model that calculates the best quadratic surface approximation of the disparity map based on the uncertain depth measures. A multi grid surface reconstruction algorithm described by [Terzopoulos,1988] was chosen to calculate the interpolation with a finite element approximation. The interpolation fills out gaps in areas where no disparity calculation is possible. The process of segmentation and model building is shown in Fig. 3 for a pair of the image sequence "house". In Fig.3a the segmentation of the object "house" is marked grey, the residual background is marked white. Black regions are areas that cannot be analyzed at all because these areas are visible in one camera only. Fig. 3b displays the interpolated disparity map of the object house that is converted into a depth map for model building.The segmentation could be improved if not only the outer



a) Scene segmentation mask

b) interpolated disparity map (dark = far from camera, light = near to camera)

c) textured object "house" with triangular surface mesh superimposed

d) Synthized view of object "house" after integration of depth measurements from 6 view points, each view rotated 4 degrees

Fig. 3: Scene segmentation and model building.

boundary of the house but also the inner edges were known. By simple human interaction these edges could be classified and would greatly improve surface modeling. A human interaction module that introduces knowledge about the objects will therefore be of great help.

Modeling of 3D–objects from the disparity map

The interpolated disparity map contains the visible scene geometry measured from the camera view point. When the scene contains occluded surfaces then the camera must be moved around the objects and the measurements from multiple view points must be included. For that purpose the 2D disparity map is first converted into a 3D surface description that can be modified to include hidden surfaces. The disparity map can be transformed into a depth map containing absolute scene geometry when the binocular camera geometry parameters are known. This depth map is sampled and a 3D–surface for each contiguous object is constructed by spanning a triangular wireframe in space. The samples of the depth map generate control points in space that are connected through plane triangular patches. The triangular mesh was chosen because it is capable to approximate arbitrary surface geometries without singularities. On the surface of each triangular patch the object surface texture can be stored in a texture map from which a naturally looking view of the original objects can be synthesized with computer graphics methods. In Fig. 3c a synthesized view of the 3D object "house" is shown with the triangular surface mesh superimposed as black lines. The surface geometry was calculated from the interpolated disparity map while the surface texture was taken from the left original image. In Fig. 3d a first result of the information fusing stage is shown which will be described below.

IMAGE SEQUENCE PROCESSING

The tasks performed so far were straight forward stereoscopic image analysis. From the stereoscopic image pair a 3D surface approximation was extracted from a single camera view point together with a quality measure of the estimated surface position. When complex scenes with occluding objects are to be analyzed or when the measurement quality of the surface geometry has to be improved the camera must be moved throughout the scene and the measurements from multiple view points have to be integrated into the 3D surface model. Therefore it is necessary to estimate the 3D motion of the camera and possible object motions in the scene from the image sequence and to fuse the multiple depth measurements into a consistent 3D scene model.

3D motion estimation using analysis by synthesis

In this section an algorithm to directly estimate 3D scene motion from a monocular image sequence is described. For 3D motion estimation the object shape is assumed to be known. An initial estimate of the scene shape was generated from stereoscopic image analysis. When the initial estimate fails this dependency may affect the analysis and will sometimes lead to estimation errors. As long as the initial shape approximation is reliable, however, this dependency can be neglected.

Requirements for 3D motion estimation An object is defined as a rigid 3D–surface in space that is spanned by a set of N control points. A set of six motion parameters is associated with each object. Object motion is defined as rotation of the object control points around the object center followed by a translation of the object center, measured between two successive image frames k and k+1. The object center **G** is the mean position vector of all N object control points. Each object control point $\mathbf{P}_{i(k)}$ at frame k is transformed to its new position $\mathbf{P}_{i(k+1)}$ in frame k+1 according to the general motion Eq. (3) between frame k and k+1.

$$\mathbf{P}_{i(k+1)} = [\mathbf{R}_G] \cdot (\mathbf{P}_{i(k)} - \mathbf{G}) + [\mathbf{R}_G] \cdot \mathbf{G} + \mathbf{T} \qquad (3)$$

$$\text{with } \mathbf{T} = \begin{pmatrix} T_x \\ T_y \\ T_z \end{pmatrix} = \begin{matrix}\text{translation} \\ \text{vector}\end{matrix} \quad ,$$

$$\mathbf{G} = \begin{pmatrix} G_x \\ G_y \\ G_z \end{pmatrix} = \sum_{i=1}^{i=N} \frac{\mathbf{P}_i}{N} = \begin{matrix}\text{component} \\ \text{center}\end{matrix} \quad , \text{ and}$$

$$[\mathbf{R}_G] = \text{rotation matrix of rotation vector } \mathbf{R} = \begin{pmatrix} R_x \\ R_y \\ R_x \end{pmatrix}$$

Object rotation can be expressed by a rotation vector $\mathbf{R} = (R_x, R_y, R_z)^T$ that describes the successive rotation of the object around the three axes $(x, y, z)^T$ parallel to the scene coordinate system centered at **G**. From this vector the rotation matrix $[R_G]$ is derived when the identical matrix [I] is rotated around the coordinate axes with R_x first, R_y second and R_z last. Because $[\mathbf{R_G}]$ is derived from the rotation vector **R**, the six parameters of **T** and **R** suffice to describe the 3D object motion.

The only information available to the analysis system is the surface texture projected onto the camera target throughout the image sequence. From this sequence the shape and motion parameters have to be derived. Assume a scene with an arbitrarily shaped, moving textured object observed by a camera **C** during frames k and k+1 as shown in Fig. 4. The object moves between frames k (dashed object silhouette) and k+1 (solid object silhouette) according to the general motion equation Eq. (3) with motion parameters **R** and **T**. A point on the object surface, called observation point $\mathbf{P}_{(k)}$, holds the surface intensity I_1, which is projected onto $\mathbf{p}_1$ in the image plane at frame k. At frame k+1 $\mathbf{P}_{(k)}$ has moved to $\mathbf{P}_{(k+1)}$, still holding I_1. In image frame k+1 the surface intensity I_1 will now be projected at $\mathbf{p}_2$, whereas the image intensity at point $\mathbf{p}_1$ has changed to I_2.

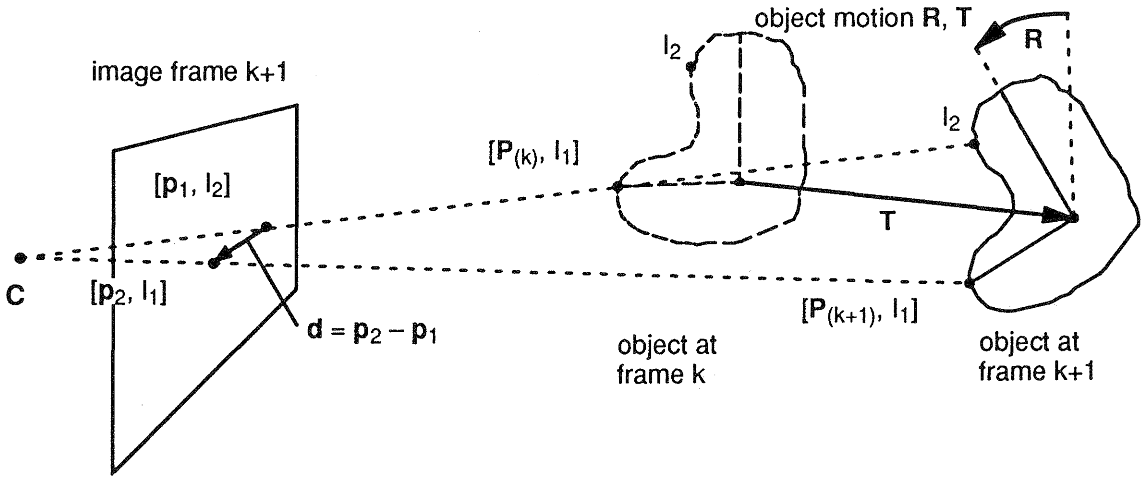


Fig. 4: Geometry for 3D motion analysis.

The image displacement vector $\mathbf{d} = \mathbf{p}_2 - \mathbf{p}_1$ is called optical flow vector and describes the projection of the observation point displacement $\mathbf{P}_{(k+1)} - \mathbf{P}_{(k)}$ onto the image plane. When assuming a linear dependency of the surface texture between I_1 and I_2 and a brightness constancy constraint between frame k and k+1 it is possible to predict I_2 from I_1 and its corresponding image intensity gradients and hence to estimate $\mathbf{d}$ from the measurable difference $I_2 - I_1$.

$$I_2 - I_1 = \left(\frac{\partial I_1}{\partial x}, \frac{\partial I_1}{\partial y}\right)^T \cdot \mathbf{d} \tag{4}$$

I_2 is measured at position of $\mathbf{p}_1$ at frame k+1, where I_1 is taken from image position $\mathbf{p}_1$ at frame k. When approximating the spatial derivatives as finite differences the optical flow vector $\mathbf{d} = (d_x, d_y)^T$ can be predicted from the image gradients $\mathbf{g} = (g_x, g_y)^T$ and the temporal image intensity difference $\Delta I_{p1} = I_2 - I_1$ between frame k+1 and k at $\mathbf{p}_1$ in Eq. (5):

$$\begin{aligned} \Delta I_{p1} &= \mathbf{g} \cdot \mathbf{d} = g_x \cdot d_x + g_y \cdot d_y \\ &= g_x \cdot (p_{2x} - p_{1x}) + g_y \cdot (p_{2y} - p_{1y}) \end{aligned} \tag{5}$$

In Eq. (5) $\mathbf{d}$ is related to intensity differences. Substituting the perspective projection of $\mathbf{P}_{(k)}$ and $\mathbf{P}_{(k+1)}$ for $\mathbf{p}_1$ and $\mathbf{p}_2$ in Eq. (5) yields a direct geometric to photometric transform that relates the spatial movement of $\mathbf{P}$ between frame k and k+1 to temporal intensity changes in the image sequence at $\mathbf{p}_1$.

$$\begin{aligned} \Delta I_{p1} &= f \cdot g_x \cdot \left(\frac{P_{(k+1)x}}{P_{(k+1)z}} - \frac{P_{(k)x}}{P_{(k)z}} \right) \\ &+ f \cdot g_y \cdot \left(\frac{P_{(k+1)y}}{P_{(k+1)z}} - \frac{P_{(k)y}}{P_{(k)z}} \right) \end{aligned} \tag{6}$$

As long as the motion between $\mathbf{P}_{(k)}$ and $\mathbf{P}_{(k+1)}$ can be expressed parametrically by a linear equation and assuming that the initial position of $\mathbf{P}_{(k)}$ is known, $\mathbf{P}_{(k+1)}$ can be substituted and the Eq. (6) can be solved through evaluation of the spatial and temporal intensity differences at $\mathbf{p}_1$. Essentially every surface point can be taken as an observation point. It is, however, useful to restrict the number of observation points to those carrying relevant information. Eq. (6) uses image intensity as well as the spatial image gradients. Image areas with zero gradient can not be used for parameter estimation. A lower bound to the image gradient is additionally introduced to account for camera measurement noise. It is therefore necessary to impose a minimum gradient threshold when selecting observation points as described by [Hötter,1988]. For each observation point its initial 3D position on the object surface, image intensity and spatial image gradients are recorded. During the analysis each observation point is projected and the intensity difference to the real image is evaluated.

Direct estimation of 3D object motion With the proposed approach, rigid 3D object motion can be estimated directly from the image sequence when the object shape is known. Assume an observation point with a known position $\mathbf{P}_{(k)}$ that moves in space and is observed by a camera $\mathbf{C}$ in Fig. 4. The motion is governed by the general motion equation (3). Assuming that rotation between successive images is small, the rotation matrix $[\mathbf{R_G}]$ can be linearized to Eq. (7).

$$[\mathbf{R}'] = \begin{bmatrix} 1 & -R_z & R_y \\ R_z & 1 & -R_x \\ -R_y & R_x & 1 \end{bmatrix} \tag{7}$$

rotation matrix $[\mathbf{R_G}]$ is linearized to $[\mathbf{R}']$

when setting $\sin\Phi \approx \Phi$ and $\cos\Phi = 1$

When substituting [R'] into Eq. (3) a linearized version for the general motion equation is found. $\mathbf{P}_{(k+1)}$ is expressed in explicit form in Eq. (8):

$$\mathbf{P}_{(k+1)} = \begin{pmatrix} P_{(k+1)x} \\ p_{(k+1)y} \\ P_{(k+1)z} \end{pmatrix} = \tag{8}$$

$$\begin{pmatrix} P_{(k)x} - (P_{(k)y} - G_y)R_z + (P_{(k)z} - G_z)R_y + T_x \\ (P_{(k)x} - G_x)R_z + P_{(k)y} - (P_{(k)z} - G_z)R_x + T_y \\ -(P_{(k)x} - G_x)R_y + (P_{(k)y} - G_y)R_x + P_{(k)z} + T_z \end{pmatrix}$$

The parameters to be estimated are the translation $\mathbf{T}$ and the rotation $\mathbf{R}$. When substituting $\mathbf{P}_{(k+1)}$ from Eq. (8) in Eq. (6) and linearizing the resulting non linear equation through Taylor expansion for $\mathbf{R}$ and $\mathbf{T}$ at $\mathbf{R} = \mathbf{T} = \mathbf{0}$, the linearized equation for a single observation point $\mathbf{P}_{(k)}$ is computed as

$$\begin{aligned} \Delta I_{p1} &= f \cdot g_x / P_z \cdot \mathbf{T_x} \\ &+ f \cdot g_y / P_z \cdot \mathbf{T_y} \\ &- f \cdot (P_x g_x + P_y g_y) / P_z^2 \cdot \mathbf{T_z} \\ &- f \cdot [P_x g_x (P_y - G_y) + P_y g_y (P_y - G_y) \\ &+ \quad P_z g_y (P_z - G_z)] / P_z^2 \cdot \mathbf{R_x} \\ &+ f \cdot [P_y g_y (P_x - G_x) + P_x g_x (P_x - G_x) \\ &+ \quad P_z g_x (P_z - G_z)] / P_z^2 \cdot \mathbf{R_y} \\ &- f \cdot [g_x (P_y - G_y) - g_y (P_x - G_x)] / P_z \cdot \mathbf{R_z} \end{aligned}$$

with $\mathbf{P}_{(k)} = (P_x, P_y, P_z)^T$. (9)

Conditions for robust motion estimation At least six distinctive observation points that lead to six linear independent equations are needed to solve for the six motion parameters $\mathbf{R}$ and $\mathbf{T}$. In real imaging situations the measurements of the spatial and temporal derivatives are noisy and some of the observation points selected may be linear dependent of each other. To cope with those conditions an over constrained set of equations is established and a linear regression is carried out using least squares fit. All observation points of one object are evaluated. It is important to note that we do **not** measure optical flow locally and then try to combine the flow field. Instead **all** observation points of a rigid surface are used to solve for $\mathbf{R}$ and $\mathbf{T}$. To account for the linearizing, the estimation is iterated. The position $\mathbf{P}$ of each observation point is initially determined by object shape and position. An estimate of the parameters **R** and **T** is calculated and the observation point is moved according to those parameters. The estimation is repeated with the new starting position of **P** until the parameter changes of **T** and **R** converge to zero.

To improve estimation stability, dependencies between rotation and translation parameters were cancelled out through the introduction of a center of rotation **G**. The rotation of an object around an arbitrary rotation center can be separated into a rotation of the object around the object's center of gravity and an additional translation of the object. Such a decomposition leads to an independent estimation of **R** and **T** and improves convergence of the solution.

The system should be robust against noisy measurements or measurements which are erroneous due to invalid model assumptions. Therefore the mean temporal intensity is computed and observation points with high intensity errors are excluded from the regression in a modified least squares fit. The measurement certainty of each parameter can be estimated through evaluation of the error covariance matrix of the regression [Hötter, 1988]. When a parameter has an uncertain estimate it can be excluded from the regression to ensure a stable estimate for the remaining parameters. The analysis was calculated from a monocular image sequence only. It has been tested successfully on a variety of tasks for object and camera motion tracking [Kappei, 1988],[Liedtke, 1990], [Welz, 1990]. When including the stereoscopic sequence information the quality of the analysis is expected to improve further.

Integration of multiple depth maps into a common 3D scene model

For each image pair of the sequence a depth map D_k can be calculated by stereoscopic analysis together with its associated confidence map C_k. The 3D scene model contains the approximated scene geometry that can be moved according to the camera and scene motion. It is now possible to fuse the depth measurements from multiple view points into the 3D scene model to improve estimation quality. The confidence value C is converted into the weight S that can easily be accumulated throughout the sequence. Each control point of the scene objects holds not only its position $\mathbf{P}_{old}$ in space but also its corresponding confidence weight S_{old}. When a new measurement becomes available, the scene motion is compensated and the new depth estimate $\mathbf{P}_{new}$ with corresponding confidence weight S_{new} is integrated by weighted accumulation. S_{fuse} represents the accumulated quality measure and $\mathbf{P}_{fuse}$ the new control point position.

$$S_{fuse} = S_{old} + S_{new} \quad \text{with } S = \frac{C}{1-C} \qquad (10)$$

and $$\mathbf{P}_{fuse} = \frac{\mathbf{P}_{old} \cdot S_{old} + \mathbf{P}_{new} \cdot S_{new}}{S_{old} + S_{new}}$$

The information fusing process described above can only be applied to an existing surface. When new objects and prior unseen object surfaces appear, the surface mesh must be extended from the new depth map. Once the surface is built, the fusing process can continue.

First results of the sequence analysis are shown in Fig. 3d with the sequence "house". The house was rotated on a turn table and 90 stereoscopic views of the house from all directions, each view displaced by 4 degree rotation, were taken. Starting with the 3D object shown in Fig. 3c, the 3D motion and rotation of the house was estimated successfully. At present the sequence analysis was tested with objects generated from a single depth map only. The object part visible from from one camera position was generated and this object part was tracked throughout the sequence, integrating the depth measurements from the different view points. The resulting object surface after integration from 6 different view points (0, 4, 8, 12, 16, and 20 degree rotation) is shown in Fig. 3d. The object is rotated to a side view to show the still existing shape deviations.

We are currently working to improve the motion estimation by fully exploiting the stereoscopic sequence information and to enhance the integration process. It is necessary that the 3D object surfaces are generated not only from a single depth map but incrementally when new surfaces appear. Additional quality measures can be thought of that govern the global surface shape and allow to introduce scene specific knowledge.

ACKNOWLEDGEMENT

This work has been supported by a grant of the German postal service TELEKOM.

REFERENCES

Aloimonos, J., Shulman, D., 1989. Integration of Visual Modules, Academic Press, San Diego, USA.

Hötter, M.,Thoma, R., 1988. Image segmentation based on object oriented mapping parameter estimation, SIGNAL PROCESSING, Vol. 15(3), pp. 315–334.

Kappei, F., 1988. Modellierung und Rekonstruktion bewegter dreidimensionaler Objekte aus einer Fernsehbildfolge, Ph.D. Thesis, University of Hannover.

R. Koch, 1990. Automatic Modelling of Natural Scenes for Generating Synthetic Movies, Eurographics '90, Montreux, Switzerland.

Liedtke, C. E., Busch, H., Koch, R., 1990. Automatic Modelling of 3D Moving Objects from a TV Image Sequence, SPIE Conf. Sensing and Reconstruction of 3D–Objects and Scenes, Vol. 1260, pp. 230–239, Santa Clara, USA.

Terzopoulos, D., 1988. The computation of visible–surface representations, IEEE Trans. Patt. Anal. Mach. Intell., Vol 10, pp.417–438.

Welz, K. 1990. Beobachtung von Verkehrszeichen aus einem bewegten Fahrzeug, Proceedings of the 7. Aachener Symposium für Signaltheorie ASST '90, Aachen, F.R.G.