3D OBJECT–BASED DEPTH ESTIMATION FROM STEREOSCOPIC IMAGE SEQUENCES

Lutz Falkenhagen

Universität Hannover
Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung
Appelstr. 9A, 30167 Hannover
phone: +49–511–762–5312, fax: +49–511–762–5333
email: falkenhagen@tnt.uni–hannover.de

ABSTRACT

A 3D object–based depth estimation approach from stereoscopic image sequences is presented in this paper that introduces a memory for depth information from preceding image pairs. This approach is used in the EU–RACE project DISTIMA for applications like remote handling, quality control and TV post processing. Compared to depth estimation approaches that compute depth maps from consecutive stereoscopic image pairs independently by a triangulation of corresponding image points, the temporal coherence and accuracy of depth estimates is increased. The remaining depth estimation error variance is reduced and as a result of temporal filtering the effect of single outliers is reduced as well.

1. INTRODUCTION

For applications like remote sensing, quality control and TV post processing, accurate depth maps from stereoscopic image sequences are needed. Common approaches estimate depth by a triangulation of corresponding image points or image features in single stereoscopic image pairs [1] [2] [3]. The search for corresponding image points or image features is affected by camera noise which causes inaccurate depth estimates. Furthermore, depth estimates are temporally incoherent, because consecutive image pairs are evaluated independently.

The basic idea of the depth estimation approach presented in this paper is to introduce a memory for depth information from preceding image pairs that reduces the temporal incoherence and increases the accuracy of depth estimation. For the use of a depth memory, the motion between consecutive images has to be estimated and compensated. [4] proposes a depth estimation approach assuming static scenes and a moving stereo camera, whose motion is estimated from the captured image sequence. In this paper a dynamic scene is considered. Assuming that the real world is composed of independently moving 3D objects, a 3D object–based approach with a 3D model of the real world is applied. This model represents each real object of the captured scene by a 3D model object that is described by its 3D shape, 3D motion and surface texture parameters [5], [6]. These three parameter sets have to be estimated from the input image sequence.

To calculate a current depth map, first a temporally independent depth map (TI depth map) is computed from the current image pair by means of correspondence analysis and triangulation. Then, the 3D motion of the model objects is estimated and compensated. From the motion compensated model objects a second depth map is computed, called motion compensated depth map (MC depth map). The current depth map is calculated by averaging both depth maps depending on the reliability of the respective depth estimates with a Kalman filter approach. The current depth map is used for an update of the 3D shape parameters of the model objects. The object–based depth estimation approach is assessed using a stereoscopic image sequence from a scene with known geometry and known stereoscopic camera setup. The results are compared with depth estimation results obtained by a common depth estimation approach.

In Chapter 2, the concept of object–based depth estimation is described. Chapter 3 explains the initialization of the 3D model objects. Chapter 4 describes the averaging of the depth maps and the update of the 3D model objects. Depth estimation results for a DISTIMA image sequence are presented in Chapter 5. In Chapter 6 the paper is concluded.

2. CONCEPT OF 3D OBJECT–BASED DEPTH ESTIMATION

The goal of the presented approach is the estimation of accurate and temporal coherent depth maps. Therefore an object–based approach has been developed that stores depth information of preceding images in an object parameter memory [7].

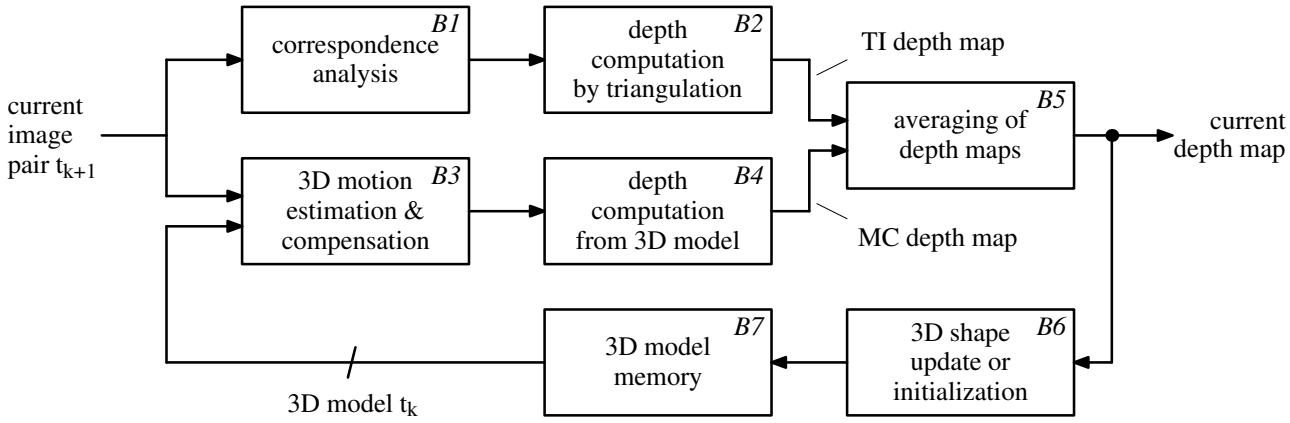


Fig. 1: Block diagram of 3D object–based depth estimation

The object–based depth estimation approach can be subdivided into several subtasks, as illustrated in Fig 1. From the current image pair, a disparity map is estimated by means of correspondence analysis (Fig. 1 *B1*). A temporally independent depth map (TI depth map) is computed from disparity by triangulating corresponding image points (Fig. 1 *B2*). For these two subtasks any depth estimation approach can be used that provides reliability measures for the estimated depth values. Gradient based approaches [3] provide the error covariance of the depth estimates as a reliability measure. If a search based approaches like [8] are applied, the error covariance of depth estimates has to be modelled using the correspondence measure, i.e. the cross–correlation of image blocks. In this paper, the scanline based approach proposed in [8] is applied, but will not further be explained.

The first additional subtask of the object–based approach is to estimate 3D motion of already modelled 3D objects from both stereoscopic images simultaneously (Fig. 1 *B3*). Then from the motion compensated 3D object a motion compensated depth map (MC depth map) is computed by projecting the distance between camera and 3D object into the image plane, a method known in computer graphics (Fig. 1 *B4*). The MC depth map is averaged with TI depth map using a Kalman filter approach (Fig. 1 *B5*). Finally the shape parameters of already existing 3D model objects are updated or new objects appearing in the scene are modelled with respect to the current depth map (Fig. 1 *B6*). The 3D model objects are stored for an evaluation of the next image pair (Fig. 1 *B7*).

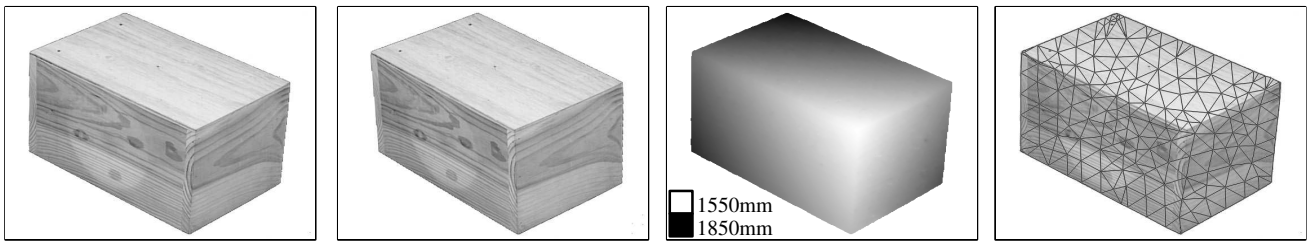
3. INITIALIZATION OF 3D MODEL OBJECTS

For object–based depth estimation, each object of the real scene is represented by a 3D model object, defined by its 3D motion, 3D shape and surface texture parameters. For the segmentation of objects it is assumed that the scene is composed of large objects of different depth and motion. When a new object appears in the scene or a new part of an object becomes visible, a new 3D model object has to be initialized. The initialization is carried out in two steps, initialization of 3D shape parameters and initialization of surface texture parameters.

3.1 Initialization of 3D Shape Parameters

The shape of the objects is represented by a 3D triangular mesh, which is generated for each segmented object. Objects are segmented by dividing the depth map into regions of homogeneous depth, assuming that object boundaries coincide with depth discontinuities. The segmentation approach is described in [8].

The boundary of each segmented region is approximated by a polygon and then filled by a 2D triangular mesh from the outside to the inside. A 3D triangular mesh is obtained by projecting the 2D triangular mesh onto the estimated depth map and determining the spatial coordinates of the vertices of the 3D triangular mesh from the depth map. An example of the resulting triangular mesh is shown in Fig 2. The size of the triangles determines the accuracy of the approximation of the depth map, because the shape of the object is assumed to be planar inside the triangles.



Left original image | Right original image | Estimated depth map | Resulting 3D model object

Fig. 2: Initial 3D model object of DISTIMA image sequence 'box'

Because of the later use of a Kalman filter, not only the spatial position of the vertices but also their reliability is stored. The reliability of depth estimates is expressed by their depth error variance σ_{de}^2 which is provided by the depth estimation approach. It is stored by assigning it as a position error covariance matrix $\underline{C}_P$ to each vertex. Beside the stochastical error of depth estimation, a deterministic error occurs by approximating the shape of the objects inside the triangles by planes. This error is known and could be compensated by storing a differential depth map for each triangle. In order to keep the complexity of the 3D model objects small, this error is modelled by a single stochastic depth error measure $\sigma_{de_T}^2$ for each triangle. $\sigma_{de_T}^2$ depends on the approximation error inside the triangle and is stored in the object–based memory.

3.2 Initialization of Surface Texture Parameters

The texture parameters of the model object are small images which are mapped onto each triangle. They are needed for image analysis purposes only. Therefore a simple estimation approach has been chosen, where the color parameters are taken from one of the images in order to keep texture consistent. No optimization yielding at an improved quality of the synthesized images has been considered.

4. AVERAGING OF DEPTH MAPS AND UPDATE OF 3D MODEL OBJECTS

4.1 Model of the Stereoscopic Camera System

In the course of image analysis, a pinhole camera model based on the central projection of diffuse illuminated and diffuse reflecting spatial points in the image plane is applied. No lense distortion is considered. Each camera is defined with respect to a Cartesian model coordinate system by its position $\vec{C}$, its optical axis $\vec{A}$ and its image plane, which is determined by two vectors $\vec{H}$ and $\vec{V}$. The camera model is therefore called CAHV camera model [9].

The projection $\vec{p}$ of a spatial point $\vec{P}_S = (P_{Sx}, P_{Sy}, P_{Sz})$ into the image plane of a CAHV–camera can be computed using the projection equation (1).

$$\vec{p} = \begin{pmatrix} p_h \\ p_v \end{pmatrix} \quad with \quad p_h = \frac{(\vec{P}_S - \vec{C}) \cdot \vec{H}}{(\vec{P}_S - \vec{C}) \cdot \vec{A}} \quad , \quad p_v = \frac{(\vec{P}_S - \vec{C}) \cdot \vec{V}}{(\vec{P}_S - \vec{C}) \cdot \vec{A}} \tag{1}$$

4.2 Estimation of 3D Motion Parameters

When a new image pair is analyzed, 3D motion of the objects has to be estimated. Therefore a gradient–based 3D motion estimator for monocular image sequences [5] was modified to be used for stereoscopic image sequences. The motion estimator assumes rigid 3D objects, whose motion can be expressed by a set of three translation parameters $\vec{T} = (T_x, T_y, T_z)$ and three rotation parameters $\vec{R} = (R_x, R_y, R_z)$. The rotation around the center of the 3D model objects $\vec{G}$ is expressed by a rotation matrix $\underline{R}$ composed of components of $\vec{R}$. The motion equation of a single object point $\vec{P}$ is given by (2).

$$\vec{P}' = \underline{R} \cdot \left(\vec{P} - \vec{G}\right) + \vec{T} + \vec{G} \tag{2}$$

A gradient based method is applied to estimate the set of motion parameters $\vec{M} = (\vec{T}, \vec{R})$. Each 3D model object is projected into the image planes and the difference to the current image signal ΔI is evaluated with the image signal gradient $\vec{g}$.

$$\Delta I = I(\vec{p}') - I(\vec{p}) = \vec{g} \cdot (\vec{p}' - \vec{p}) \quad with \quad \vec{g} = \begin{pmatrix} g_h \\ g_v \end{pmatrix} \tag{3}$$

The estimator requires a linear dependence between the estimation parameters $\vec{M}$and the measurements ΔI. Therefore the image coordinates $\vec{p}$ are transformed into spatial coordinates $\vec{P}$ (1) and coordinates after motion $\vec{P}'$ are expressed by coordinates before motion $\vec{P}$ and the motion parameters $\vec{M}$ (2). The resulting equation is then linearized leading to a scalar product between $\vec{M}$ and a set of coefficients $\vec{\Phi}$:

$$\Delta I = \vec{\Phi}^T \cdot \vec{M} \tag{4}$$

Applying (4) to a large number of object points, so–called observation points, enables the estimation of $\vec{M}$ by using the Gaussian–Estimator. The Gaussian–Estimator assumes that all measured luminance differences are uncorrelated and have the same error variance. When the measurement error covariance matrix $\underline{C}_{\Delta I}$ is known, the Gaussian–Markov–Estimator can be applied that minimizes and determines the estimation error variance of the motion parameters $\underline{C}_M$. Because $\underline{C}_M$ is required for the computation of the position error covariances of the vertices after motion compensation, the Gaussian–Markov–Estimator is used.

$$\vec{M} = (\underline{H}^T \cdot \underline{C}_{\Delta I}^{-1} \cdot \underline{H})^{-1} \cdot \underline{H}^T \cdot \underline{C}_{\Delta I}^{-1} \cdot \vec{\Delta I} = \underline{C}_M \cdot \underline{H}^T \cdot \underline{C}_{\Delta I}^{-1} \cdot \vec{\Delta I} \tag{5}$$

Assuming that the luminance differences of the observation points are uncorrelated, only the luminance error variance $\sigma_{\Delta I}^2$ has to be determined for using the Gaussian–Markov–Estimator. $\sigma_{\Delta I}^2$ is depends on the known camera noise and on position errors of the observation points that lead to luminance differences. The luminance error variance $\sigma_{\Delta I}^2$ resulting from a position error variance $\underline{C}_P$ is modelled by projecting $\underline{C}_P$ onto the image gradient. After motion compensation, $\underline{C}_P$ is increased due to motion estimation failures, which is considered using:

$$\underline{C}_P{}' = \underline{R} \cdot \underline{C}_P \cdot \underline{R}^T + \underline{C}_M \tag{6}$$

4.3 Averaging of Depth Maps

The accuracy of depth estimation can be increased by averaging corresponding depth values of the two independently estimated depth maps, if the reliability of the depth estimates is considered. Therefore the depth value of the TI depth map de_{TI} and the motion compensated depth map de_{MC} are averaged with respect to their depth error variances $\sigma_{de_{TI}}^2$ and $\sigma_{de_{MC}}^2$. The depth error variance of the TI depth map $\sigma_{de_{TI}}^2$ is given by the depth estimation approach. The depth error variance of the motion compensated 3D model objects $\sigma_{de_{MC}}^2$ is the projection of position covariance of the model object point onto its current viewing line. The following Kalman filter approach is used for averaging the depth values:

$$de_{cur} = \frac{de_{MC} \cdot \sigma_{de_{TI}}^2 + de_{TI} \cdot \sigma_{de_{MC}}^2}{\sigma_{de_{MC}}^2 + \sigma_{de_{TI}}^2} \quad , \quad \sigma_{de_{cur}} = \frac{\sigma_{de_{TI}}^2 \cdot \sigma_{de_{MC}}^2}{\sigma_{de_{TI}}^2 + \sigma_{de_{MC}}^2} \tag{7}$$

4.4 Update of 3D Shape Parameters

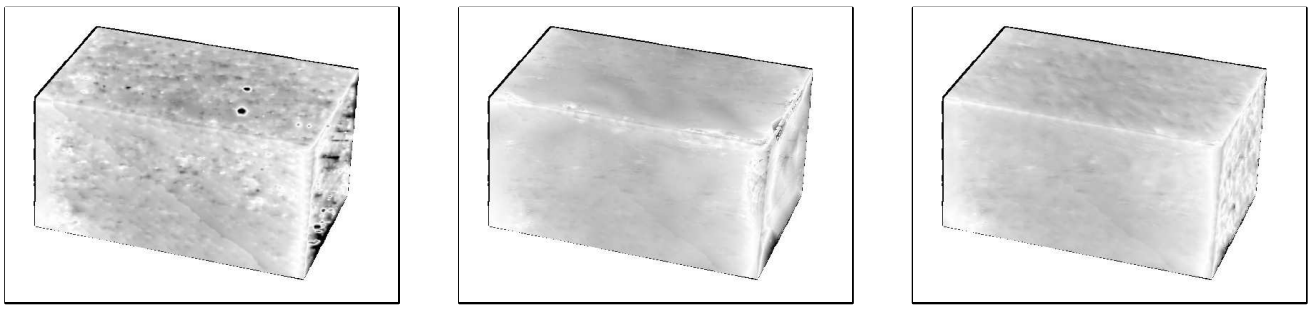
The 3D shape parameters of the model objects have to be updated after averaging the depth maps in order to keep the current depth map and model objects consistent. This update is carried out by shifting the vertices along their viewing lines according to the current depth value. The position error covariances are updated respectively.

5. DEPTH ESTIMATION RESULTS

In order to verify the estimation results, a scene with known geometry and motion was recorded. This knowledge enables the generation of ground truth reference depth maps which are used for an assessment of the estimated depth maps. The selected scene is composed of a rectangular box rotating in front of a calibrated stereoscopic camera system. The rotation is carried out on a highly precise turntable. The distance between the camera and the box is approximately 1.5m. The base length of the stereoscopic camera is approximately 0.18m. With the described arrangement, a stereoscopic image sequence of 72 image pairs has been captured. The rotation of the box between two image pairs is 5°. A set of 72 reference depth maps has been computed.

Based on the knowledge of the real motion parameters, the estimated 3D motion parameters of the model object are verified. It turned out that the rotation angle between two consecutive image pairs of 5° is estimated with an estimation error standard deviation below 0.02°. With this motion estimation accuracy, the additional depth error due to motion errors of the model object is rather small compared to the other failures. Nevertheless the error is considered while averaging the depth values.

The shape of the real objects is approximated by a triangular mesh. Of course, large triangles involve large approximation errors and vice versa. In order to investigate the influence of the approximation error, two triangle sizes have been used, resulting in two different model objects called box_{coarse} and box_{fine}. Objects of both triangle sizes have been applied for depth estimation. The resulting depth estimation error has been compared to the depth estimation error that occurs without the object–based approach. Exemplary depth error maps are shown in Fig. 3.



Depth error temporal independent | Depth error object–based box_{coarse} | Depth error object–based box_{fine}

Fig. 3: Exemplary depth estimation results of image sequence ’box’ applying different depth estimation approaches □ = 0 mm □ = 4 mm ■ = 12 mm ■ = 25 mm

A comparison of the resulting depth maps shows that there are similar depth estimation errors of about 4 mm in most areas. Object–based depth estimation gains obviously in those regions, where the temporal independent approach shows outliers. These errors result from ambiguous depth estimation in a single image pair that is prevented by the depth memory of the object–based approach.

Tab. 5.1 shows depth estimation results for image sequence ’box’. The errors are computed by comparing the depth estimates to depth values of reference depth maps. The resulting depth error of the object–based approach is slightly below the error of the temporal independent approach. The remaining depth error variance is reduced about 9% when the fine surface approximation is applied. Even more important is that as a result of temporal filtering the effect of single outliers is also reduced.

	Temporal independent	Object–based box_{coarse}	Object–based box_{fine}
Depth error variance	32,0 mm^2	31,4 mm^2	29,1 mm^2
Mean relative depth error	0,38 %	0,37 %	0.36 %

Tab. 5.1: Mean depth estimation errors applying different depth estimation approaches for image sequence ’box’ compared to a set of reference depth maps

6. CONCLUSIONS

In this paper an object–based depth estimation approach has been presented. This approach calculates a current depth map for each stereoscopic image pair by suitably averaging two independently estimated depth maps, the temporally independent depth map (TI depth map) and the motion compensated depth map (MC depth map). The TI depth map is computed using a scanline–based approach [8]. The MC depth map is computed from the motion compensated 3D model objects.

When a new object or object part occurs, the shape and surface texture of the representing 3D model object are initialized. The shape is estimated by segmenting the TI depth map into regions of homogeneous depth and treating each segmented region as the silhouette of an independently moving object. Each region is approximated by a triangular mesh with spatial coordinates taken from the TI depth map. The position error variance of the spatial coordinates, which is provided by the TI depth estimation approach, is stored together with the shape parameters of the 3D model object. For texture initialization the color of the left image is mapped onto corresponding triangles of the 3D model object.

The motion of the objects is estimated from the current image pair simultaneously using a gradient based approach. The error variance of the estimated motion parameters is determined by a Gaussian–Markov–Estimator. The motion error variance increases the position error variance of the shape parameters, which is considered during motion compensation.

Finally, the current depth map is computed by averaging the TI and the MC depth map. The averaging of the depth values is carried out considering their error variances using a Kalman filter approach. Assuming that the estimated depth values of both depth maps are statistically independent and that their errors are Gaussian distributed, the position error of the resulting depth values is reduced. The current depth map and position error variances are used for an update of the shape parameters of the 3D model object.

In order to assess the motion and depth estimation results, the image sequence 'box' has been used, where the geometry of the scene, the camera arrangement and the motion parameters are known. Based on this knowledge, a set of ground truth reference depth maps has been computed. It turned out that the motion of the object, which is a rotation of 5° between consecutive image pairs, can be estimated with a remaining error below 0.02°. Due to the relatively accurate motion estimation, the position error variance of the shape parameters caused by motion estimation failures is small compared to position error variances caused by depth estimation failures. Depth maps resulting from the scanline–based approach and the object–based approach have been compared to the reference depth maps. Applying the object–based approach reduces the remaining depth error variance from 32.0 mm^2 to 29.1 mm^2. This depth error corresponds to a mean error of 0.36 % referring to the distance between the camera and the object. Even more important is that as a result of temporal filtering the effect of single outliers is also reduced.

Beside the usage in DISTIMA project applications, the presented approach has also been used for 3D–modelling of natural scenes in the field of architecture visualization and mixing of natural and synthetic scenes in TV production (RACE MONALISA project). It will also be used as a basis for 3D modelling approaches in the ACTS PANORAMA project that aims at the development of a communication system with increased telepresence.

7. ACKNOWLEDGEMENT

This work was supported by the EU–RACE II project R2045 "DISTIMA".

8. REFERENCES

[1] M. J. P. M. Lemmens, "A survey on stereo matching techniques", International Archives of Photogrammetry and Remote Sensing, Vol. 27, Comm. V., pp. 11–23, 1988.

[2] A. Koschan, "Eine Methodenbank zur Evaluierung von Stereo–Vision–Verfahren", Ph. D Thesis, Technische Universität Berlin, Berlin, Germany, 1991.

[3] W. Förstner, "A feature based corresponding algorithm for image matching", International Archives of Photogrammetry and Remote Sensing, Vol26–3, 1986.

[4] A. P. Tirumali, et.al, "Dynamic Stereo with Self–Calibration", IEEE Transactions on PAMI, Vol. 14, No.12, 1992.

[5] H. G. Musmann, M. Hötter and J. Ostermann, "Object–Oriented Analysis–Synthesis Coding of Moving Images", Signal processing: Image communication, Vol 1, No. 2, October 1989, pp. 117 – 138.

[6] R. Koch, "Dynamic 3–D Scene Analysis through Synthesis Feedback Control", IEEE Transactions on PAMI, Vol. 12, No. 6, June 1993

[7] R. Koch, "3–D Surface Reconstruction from Stereoscopic Image Sequences", International Conference of Computer Vision ICCV '95, Cambridge, MA., USA, June 1995.

[8] L. Falkenhagen, "Depth Estimation from Stereoscopic Image Pairs Assuming Piecewise Continuos Surfaces", in: Paker, Y. and Wilbur, S. (Ed.), Image Processing for Broadcast and Video Production, Hamburg 1994, pp. 115–127, Springer series on Workshops in Computing, ISBN 3–540–19947–0, Springer Great Britain, 1994.

[9] Y. Yakimovski, R. Cunningham, "A System for Extracting 3D Measurements from a Stereo Pair of TV Cameras", CGVIP, Vol. 7, 1978, pp. 195 – 210.