3–D Scene Modeling from Stereoscopic Image Sequences

Reinhard Koch
Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung
Division "Automatic Image Interpretation", Chair Prof. C.–E. Liedtke
Universität Hannover, Appelstrasse 9A, 30167 Hannover, Germany
email: koch@tnt.uni–hannover.de

Abstract

A vision–based 3–D scene analysis system is described that is capable to model complex real–world scenes like buildings automatically from stereoscopic image pairs. Input to the system is a sequence of stereoscopic images taken with two standard CCD Cameras and TV lenses. The relative orientation of both cameras to each other is estimated by calibration. The camera pair is then moved throughout the scene and a long sequence of closely spaced views is recorded. Each of the stereoscopic image pairs is rectified and a dense map of 3–D surface points is obtained by area correlation, object segmentation, interpolation, and triangulation. 3–D camera motion relative to the scene coordinate system is tracked directly from the image sequence which allows to fuse 3–D surface measurements from different view points into a consistent 3–D scene model. The surface geometry of each scene object is approximated by a triangular surface mesh which stores the surface texture in a texture map. From the textured 3–D models, realistic looking image sequences from arbitrary view points can be synthesized using computer graphics.

1 Introduction

The rapid progress in the development of powerful computer graphics hardware and software enables users in a wide range of applications to gain a better insight into processes by visual simulation. Suppliers of flight and driving simulators as well as landscape and city planners are interested to simulate photo–realistic views of the environment. Architects and city planners for example construct new buildings with CAD systems and are interested to visualize their impact onto the existing environment beforehand. Complete realism, however, is possible only if the buildings to be constructed are placed inside a 3–D reconstruction of of the real environment. It is therefore necessary to reconstruct the existing environment as a 3–D model of the real scene with as little effort as possible [1]. One possible approach is to obtain a complete 3–D scene description by evaluating images of the scene.

Automatic evaluation of all scene properties, camera position and 3–D object geometry as well as photometric surface mapping, for the purpose to reconstruct 3–D scene models for visualization, are discussed in this contribution. To overcome the problem of simultaneous estimation of object geometry and camera position, a calibrated stereoscopic image sequence is recorded. From each image pair the geometry is measured and from the sequence information relative camera motion can be extracted. All measurements obtained from the image sequence need then to be integrated into a consistent 3–D scene model that contains not only the scene geometry but also texture maps of the object surface. Visual simulations of the scene from this complete scene model can be performed using computer graphics methods [2].

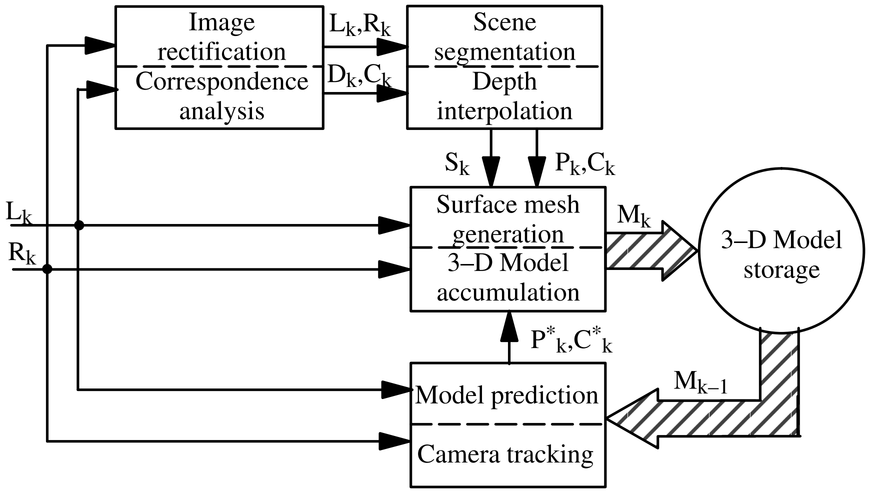
This approach is mostly data driven and models the 3–D scene directly from the recorded image sequence in a bottom–up process. It can be complemented by the model–driven approach by Grau and Tönjes [3] that uses high level semantic information about the scene contents to constrain the modeling problem in a top–down process.

The paper is organized as follows. Chapter 2 discusses the concept of the scene analysis system. Chapter 3 treats the measurement of object geometry from a single image pair whereas in Chapter 4 motion estimation and sequence accumulation is discussed. Chapter 5 concludes with some results of scene reconstruction.

2 Concept of 3–D Scene Modeling

The structure of the scene analysis process is shown in Fig. 1. The upper part consists of the image analysis pipeline that computes a model scene M_k from a stereoscopic image pair L_k, R_k at time instant k. The pipeline computes a complete surface model for the current view point k through image rectification, correspondence analysis, scene segmentation, depth interpolation, and surface mesh generation. The model is represented by a triangular surface mesh that stores the surface geometry in its control points **P** and the color from the input images as texture maps on the triangle surfaces.

Sequence information is included into the analysis pipeline by camera motion tracking and 3–D model accumulation. The camera motion is computed directly from the



L_k: Left image D_k: Disparity map S_k: Segmentation map
R_k: Right image C_k: Confidence map P_k: Depth map M_k: 3–D Scene model
k: Index indicating the actual time frame $^*k$ = data predicted from model M_{k-1}

Fig. 1: Structure of 3–D scene analysis from stereoscopic sequences.

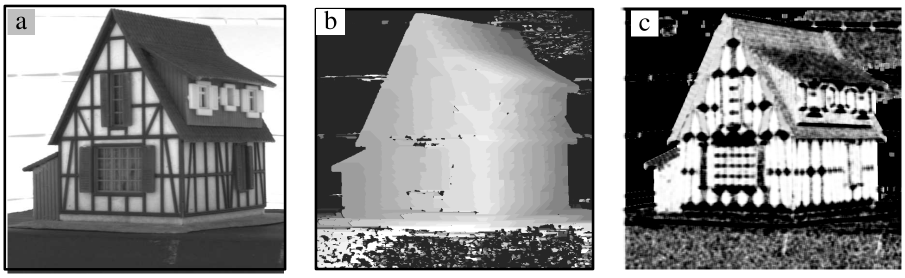
spatio–temporal image intensity gradients. Once the new camera position is known, the geometry of the existing scene model M_{k-1} can be predicted into the current camera position at frame k and merged with the new depth measurements to yield a refined scene model M_k. Merging is performed by computing an optimal estimate from the predicted surface geometry $\mathbf{P}^*_k$ and the newly measured geometry $\mathbf{P}_k$ with a Kalman filter for each control point $\mathbf{P}$.

3 3–D Surface Modeling from an Image Pair

Stereoscopic image analysis allows to compute a 3–D surface model of the scene for each image pair through correspondence analysis. For each pixel in the left image the corresponding pixel in the right image is searched for and stored in the disparity map D_k. As quality measure for the best match the cross correlation of a small block circumscribing the pixel is evaluated in combination with dynamic programming to search for the optimum disparity. The quality of the match and therefore the quality of the disparity value is recorded in a confidence map C_k. For a detailed description of the correspondence analysis please refer to the literature [4],[5]. As an example of disparity estimation the correspondence analysis for the scene HOUSE is shown in Fig. 2. Fig. 2a contains the left image of a stereoscopic image pair taken from the scene and Fig. 2b and 2c display the disparity and confidence map computed from the scene.

The correspondence analysis yields a disparity map based on local depth measurement only. These measurements are corrupted by noise and must be merged to regions that describe physical object surfaces. Based on similarity measures the segmentation divides the viewed scene into smooth object surfaces. As similarity measure the estimated disparities as well as grey level statistics are used to group pixels into regions of similar surface orientation. The region boundaries are then 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.

In order to obtain an efficient 3–D surface description and to treat hidden surfaces properly, the scene depth is represented by a triangular surface mesh. For each surface the depth map is approximated by triangular, planar surface patches. The triangular mesh was chosen because it is capable to approximate arbitrary surface geometries

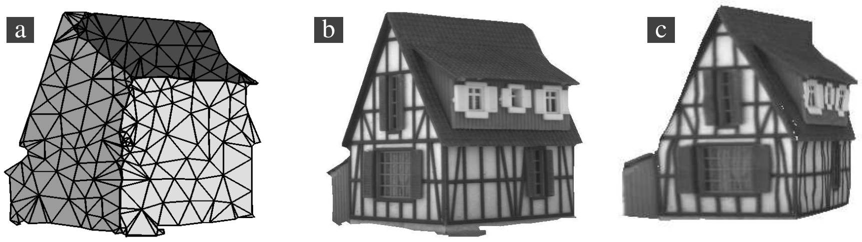


left original image

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

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

Fig. 2: Stereoscopic disparity analysis of image pair "HOUSE".



Triangular surface mesh of walls and roof | texture mapping of scene texture | synthesized image from new view point

Fig. 3: Triangulation, texture mapping and image synthesis.

without singularities. The geometric surface model computed for the scene HOUSE is shown in Fig 3a with each surface having another color. On the surface of each triangular patch the original texture is stored in a texture map from which naturally looking views can be synthesized through texture mapping, as shown in Fig. 3b for the original and in3c for a new view point.

4 Image Sequence Fusion

For each image pair of the sequence a disparity map D_k is calculated independently by stereoscopic analysis together with its associated confidence map C_k. The model geometry can be enhanced when multiple views of the scene are processed and fused together. The goal of sequence accumulation is therefore to fuse the depth measurements from the image sequence into a consistent 3–D scene model to improve estimation quality. This implies the need to estimate the camera positions for all view points.

4.1 Estimation of 3–D Camera View Point

The camera position can be derived directly from the spatial and temporal image gradients as long as the relative camera motion is small between consecutive image frames [6], [7]. It is computed by tracking the relative motion of the 3–D objects visible to the camera and then adjusting the camera position accordingly.

An object is defined as a rigid 3–D surface in space that is spanned by a set of N control points. Six motion parameters are associated with the 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–1 and k. The object center **G** is the mean position vector of all N object control points. Each object control point $\mathbf{P}_{\mathbf{i}(k-1)}$ at frame k–1 is transformed to its new position $\mathbf{P}_{\mathbf{i}(k)}$ in frame k according to the general motion Eq. (1) between frame k–1 and k.

$$\mathbf{P}_{\mathbf{i}(k)} = \underline{R} \cdot (\mathbf{P}_{\mathbf{i}(k-1)} - \mathbf{G}) + \mathbf{G} + \mathbf{T} \tag{1}$$

$$\mathbf{T} = (T_x, T_y, T_z)^T = \text{translation vector}, \quad \mathbf{G} = \sum_{i=1}^{N} \frac{\mathbf{P_i}}{N} = \text{component center}$$

$$\underline{R} = \text{matrix of rotation vector } \mathbf{R} = (R_x, R_y, R_z)^T$$

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 motion parameters have to be derived. Assume a scene with an arbitrarily shaped, moving textured object observed by a camera during frames k–1 and k. The object moves between frame k–1 and k according to the general motion Eq. (1) with motion parameters **R** and **T**. A control point on the object surface $\mathbf{P}_{(k-1)}$ holds the surface intensity I_1, which is projected onto $\mathbf{p}_1$ in the image plane at frame k–1. $\mathbf{P}_{(k-1)}$ is moved to $\mathbf{P}_{(k)}$, still holding I_1 that is now projected onto $\mathbf{p}_2$ in image frame k. In image frame k the surface intensity I_1 is projected at image position $\mathbf{p}_2$, whereas the image intensity at point $\mathbf{p}_1$ has changed to I_2.

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 $\mathbf{P}_{(k)} - \mathbf{P}_{(k-1)}$ 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–1 and k it is possible to predict I_2 from I_1 and its corresponding image intensity gradients and hence to estimate **d** from the measurable difference $I_2 - I_1$. I_2 is measured at position of $\mathbf{p}_1$ at frame k, where I_1 is taken from image position $\mathbf{p}_1$ at frame k–1. When approximating the spatial derivatives as finite differences the optical flow vector $\mathbf{d} = (d_x, d_y)^T$ can be predicted from the spatial image gradients $\mathbf{g} = (g_x, g_y)^T$ and the temporal image intensity difference $\Delta I_{\mathbf{p}1} = I_2 - I_1$ between frame k and k–1 at $\mathbf{p}_1$ in Eq. (2):

$$\Delta I_{\boldsymbol{p}1} = \boldsymbol{g}^T \cdot \boldsymbol{d} = g_x \cdot d_x + g_y \cdot d_y = g_x \cdot (p_{2x} - p_{1x}) + g_y \cdot (p_{2y} - p_{1y}) \qquad (2)$$

In Eq. (2) **d** is related to intensity differences. Substituting the perspective projection of $\mathbf{P}_{(k-1)}$ and $\mathbf{P}_{(k)}$ for $\mathbf{p}_1$ and $\mathbf{p}_2$ in Eq. (2) yields a direct geometric to photometric transform that relates the spatial movement of **P** between frame k–1 and k to temporal intensity changes in the image sequence at $\mathbf{p}_1$.

$$\Delta I_{\boldsymbol{p}1} = f \cdot g_x \cdot \left(\frac{P_{(k)x}}{P_{(k)z}} - \frac{P_{(k-1)x}}{P_{(k-1)z}} \right) + f \cdot g_y \cdot \left(\frac{P_{(k)y}}{P_{(k)z}} - \frac{P_{(k-1)y}}{P_{(k-1)z}} \right) \qquad (3)$$

With this approach, rigid 3–D object motion can be estimated directly from the image sequence when the object shape $\mathbf{P}_{(k-1)}$ is known. Assuming that rotation between successive images is small, $\underline{R}$ can be linearized and $\mathbf{P}_{(k)}$ is substituted in Eq. (3) as a function of the unknown parameter **R** and **T** as derived in Eq. (1) :

$$\begin{aligned}
\Delta I_{p1} \cdot P_z^2 \quad &= f \cdot g_x \cdot P_z \cdot \boldsymbol{T_x} + f \cdot g_y \cdot P_z \cdot \boldsymbol{T_y} - [\, \Delta I_{p1} \cdot P_z + f \cdot P_x g_x + f \cdot P_y g_y \,] \cdot \boldsymbol{T_z} \\
&- [\Delta I_{p1} \cdot P_z \cdot (P_y - G_y) + f \cdot P_x \cdot g_x \cdot (P_y - G_y) + f \cdot P_y \cdot g_y \cdot (P_y - G_y) + f \cdot P_z \cdot g_y \cdot (P_z - G_z) \,] \cdot \boldsymbol{R_x} \\
&+ [\, \Delta I_{p1} \cdot P_z \cdot (P_x - G_x) + f \cdot P_x \cdot g_x \cdot (P_x - G_x) + f \cdot P_y \cdot g_y \cdot (P_x - G_x) + f \cdot P_z \cdot g_x \cdot (P_z - G_z) \,] \cdot \boldsymbol{R_y} \\
&+ [\, f \cdot P_x \cdot g_y \cdot (P_x - G_x) - f \cdot P_z \cdot g_x \cdot (P_y - G_y)] \cdot \boldsymbol{R_z} \\
&\text{with } (P_x, P_y, P_z)^T = \boldsymbol{P}_{(k-1)}. \qquad (4)
\end{aligned}$$

Eq. (4) contains a linear dependency of the motion parameter $\mathbf{X} = (\mathbf{T},\mathbf{R})^T$ with the image gradients ΔI and **g** and can be solved when many independent control points $\mathbf{P}_i$ are evaluated using linear regression. Solving the system of equation (4) for all observation points simultaneous is accomplished by minimizing the squared error of intensity differences ΔI. The confidence of the motion estimation can be computed as motion parameter covariance $\underline{C}_X$ from the solution as well.

4.2 Fusing Depth Measurements from Multiple View Points

Depth measurements are improved by weighted depth accumulation from the motion compensated sequence of depth maps. An optimum estimate $\hat{\boldsymbol{P}}_k$ and covariance $\underline{\hat{C}_k}$ of the model control points is computed throughout the sequence by applying a Kalman filter to each control point of the surface. The Kalman filter exists of three phases: measurement, prediction, and update. Details can be found in the literature [8],[9].

Measurement: For each control point of the model surface at the current frame k a depth measurement $\mathbf{P}_k$ is computed from the analysis pipeline together with a confidence value C that expresses the measurement accuracy. A point covariance $\underline{C_{Pk}}$ can be computed from the confidence value C and the camera uncertainty $\underline{C_X}$.

Prediction: A Prediction $\mathbf{P}^*_k$ of the new point position is made based on the motion model of the objects as defined in Eq. (5) from the old control point $\mathbf{P}_{k-1}$. The point covariance $\underline{C^*_{Pk}}$ is predicted as well by the motion model in Eq. (6). This prediction propagates the existing model control points from M_{k-1} to the current frame k.

$$\boldsymbol{P}^*_k = \underline{R} \cdot (\hat{\boldsymbol{P}}_{k-1} - \boldsymbol{G}) + \boldsymbol{G} + \boldsymbol{T} \tag{5}$$

$$\underline{C^*_{P_k}} = \underline{R} \cdot \underline{\hat{C}_{P_{k-1}}} \cdot \underline{R}^T \tag{6}$$

Update: The new control point measurement is fused with the predicted old control point position by the Kalman filter. The Kalman gain $\underline{K}$ is computed from the covariance matrices and the optimum new control point position and covariance matrix is derived in Eq. (8) and (9).

$$\underline{K_k} = \underline{C^*_{P_k}} \cdot \left(\underline{C^*_{P_k}} + \underline{C_{P_k}}\right)^{-1} \tag{7}$$

$$\hat{\boldsymbol{P}}_k = \boldsymbol{P}^*_k + \underline{K_k} \cdot (\boldsymbol{P}_k - \boldsymbol{P}^*_k) \tag{8}$$

$$\underline{\hat{C}_{P_k}} = \underline{C^*_{P_k}} - \underline{K_k} \cdot \underline{C^*_{P_k}} \tag{9}$$

Fig. 4 shows the result of sequence accumulation for the scene HOUSE. A 3–D model of the house was generated and images were synthesized from a new view point by rotating the camera to 45 degrees from above. Fig. 4a displays the result of modeling from a single image pair. The model shape appears rough due to measurement errors. In Fig. 4b the result of image fusion from five different view points is shown. The model geometry is much smoother because of the temporal filtering. This result can still be improved when spatial interpolation is used to smooth the measurements as shown in Fig. 4c.

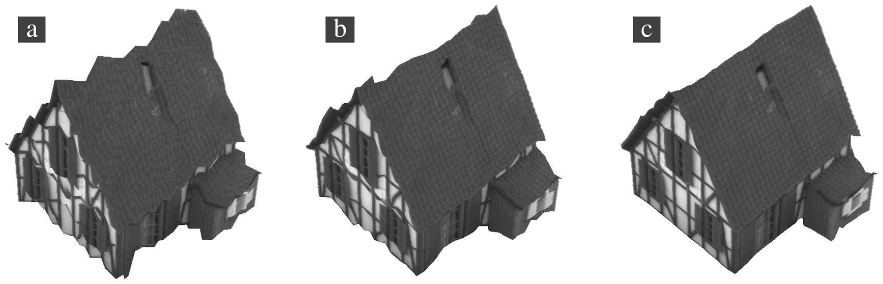


Image pair modeling   Depth fusion from 5 views   Spatio–temporal interpolation

Fig. 4: Results of temporal and spatial filtering (synthesized views).

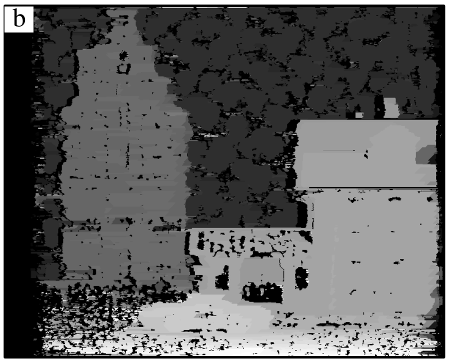
5 Conclusion

A system for automatic 3–D scene analysis was discussed. The system is capable to analyze a complex real scene from an arbitrarily moving stereoscopic video camera system. It segments the scene into smooth surfaces and stores the 3–D geometry of the scene in a 3–D scene model, including surface texture. Camera motion is tracked throughout the sequence and measurements from different view points are integrated to improve the model geometry.

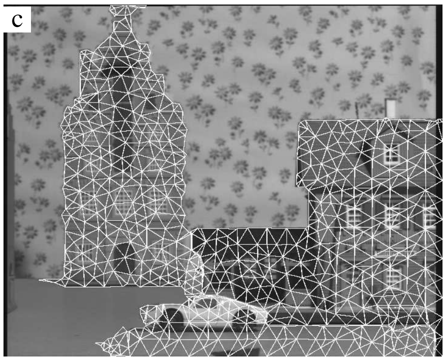
The system was tested with a variety of scenes. An example of a more complex scene with many objects is the scene STREET in Fig.5a. The scene depth was computed (Fig.5b), the scene was segmented into different objects (Fig.5c) and a 3–D model of the main scene objects are was generated and synthesized (Fig.5d).

Applications of the system are expected in many areas like automatic model generation for driving simulators, architecture visualization, or realistic image synthesis for computer generated television production.

Left original image



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



Triangular mesh of main scene objects, superimposed onto the left image

Synthesis of the main scene objects from the 3–D surface model

Fig. 5: 3–D Modeling of image pair STREET.

Acknowledgement

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

References

[1] Durisch, P. Photogrammetry and Computer Graphics for Visual Impact Analysis in Architectur. Proceedings of ISPRS Conference 1992, Vol. 29, B5, pp. 434–445.

[2] Koch, R. Automatic Modelling of Natural Scenes for Generating Synthetic Movies. In: Vandoni, C.E. and Duce, D.A. (ed.) Eurographics Association 1990. Elsevier Science Publishers B.V. (North–Holland).

[3] Grau, O., Tönjes, R. Knowledge Based Modelling of Natural Scenes. European Workshop on Combined real and synthetic image processing for broadcast and video productions, 23–24. 11. 1994, Hamburg, Germany.

[4] Falkenhagen, L. Depth Estimation from Stereoscopic Image Pairs Assuming Piecewise Continuos Surfaces. European Workshop on Combined real and synthetic image processing for broadcast and video productions, 23–24. 11. 1994, Hamburg, Germany.

[5] Cox, I., Hingorani, S., Maggs, B., Rao, S. Stereo without Regularisation. British Machine Vision Conference, Leeds, UK, David Hogg & Roger Boyle (ed.), Springer Verlag, 1992, pp. 337–346.

[6] Koch, R. Automatic Reconstruction of Buildings from Stereoscopic Image Sequences. Eurographics ’93, Barcelona,Spain, 1993.

[7] Koch, R. Dynamic 3D Scene Analysis through Synthesis Feedback Control. IEEE Trans. Patt. Anal. Mach. Intell., Special issue on analysis and synthesis 1993. Vol. 15(6):556–568.

[8] Brammer, K, Siffling, G. Stochastische Grundlagen des Kalman–Bucy–Filters. R. Oldenbourg Verlag, München, 1985.

[9] Brammer, K, Siffling, G. Kalman–Bucy–Filter: Deterministische Beobachtung und stochastische Filterung. R. Oldenbourg Verlag, München, 1985.