BLOCK–BASED DEPTH ESTIMATION FROM IMAGE TRIPLES WITH UNRESTRICTED CAMERA SETUP

L. Falkenhagen
Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung
Universität Hannover
Hannover, Germany

Abstract – A block–based depth estimation approach is presented that evaluates image triples captured with an unrestricted calibrated camera setup. From each image triple, three image pairs are formed which are evaluated independently after respective rectification. The two resulting depth maps per image, one with respect to each of the other images, are averaged according to their reliability. Compared to binocular depth estimation, the presented trinocular approach reduces the depth error variance by 40 % and increases the number of reliable depth estimates significantly. In the European ACTS PANORAMA project, it is successfully applied to 3–D reconstruction of moving objects for video communications with 3–D telepresence.

INTRODUCTION

Future video communication systems will enable a highly effective interpersonal information exchange by introducing 3–D telepresence. The European project ACTS PANORAMA develops a 3–D telepresence system, where an interactive stereoscopic visualisation of the scene with look–around capability provides the feeling of physically being close to the communication partner. The interactive visualisation requires a 3–D reconstruction of the scene that is realised using stereo vision.

Existing stereo vision approaches can be classified into two classes [1]: feature–based and area–based approaches. An area–based approach is used here, because the 3–D reconstruction requires depth information for each picture element (pel). In particular, block–matching is applied, where a correspondence analysis is carried out on squared blocks of pels. Due to geometric constraints, the search area of block–matching is reduced to the epipolar line, i.e. to the projection of a line of sight of an image point into the other image plane. From the distance between corresponding pels in both image planes, called disparity, depth values are computed.

Binocular block–based depth estimation fails in regions that are occluded in one of the images and in regions, where the luminance structure is parallel to the epipolar line.

With a trinocular camera setup as shown in Fig. 1, a third camera image is available for depth estimation. The third camera reduces the size of occluded areas and enables an evaluation of any luminance structure, because the two resulting epipolar lines are not parallel. Furthermore, the accuracy of depth estimation can be increased by using an additional camera. To allow an unrestricted camera setup, e.g. with converging optical axes on the communicating person, a respective image rectification is performed. Thus, in this paper for the first time, a trinocular block–based depth estimation approach with unrestricted camera setup is presented.

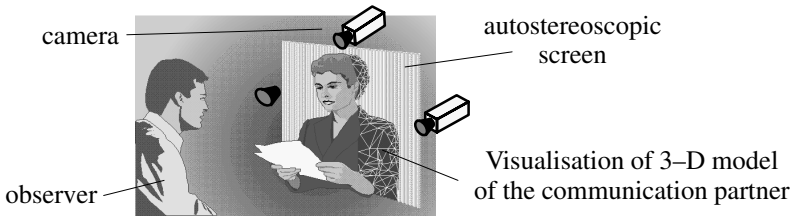


Fig. 1 Trinocular camera setup for video communications

TRINOCULAR DEPTH ESTIMATION

An optimal estimation approach would evaluate the three images simultaneously. However, with an unrestricted camera setup this would lead to an inacceptably high computational effort. Therefore, a sub–optimal approach has been developed that processes the three image pairs independently, which can be formed from an image triple. This approach is composed of the following modules:

- Binocular image rectification
- Binocular depth estimation
- Back–projection and averaging of depth estimates

Each of these modules is described in one of the following chapters.

BINOCULAR IMAGE RECTIFICATION

A rectification of each binocular image pair is applied so that epipolar lines coincide with the scan lines of the other images. This enables an unrestricted setup of the cameras together with an efficient block–matching approach. The developed rectification technique requires a calibration of the camera triple. A calibration technique derived by Tsai [2] is used that describes the camera mathematically using the CAHV camera model [3]. The CAHV camera model describes a camera by the focal point $\vec{C}$, the optical axis $\vec{A}$ and the two vectors of the image plane $\vec{H}$ and $\vec{V}$ (Fig 2).

From the real camera parameters, a rectified binocular CAHV camera parameter set is computed. Rectification has to maintain the focal points $\vec{C}_L$ and $\vec{C}_R$ in order to allow an image rectification without knowing the depth of the scene. The other three

vectors form an orthogonal coordinate system that is identical for both rectified cameras and adapted to the real camera setup [4] using (1).

$$\vec{A}_{l,r_p} = \vec{A}_{l,r} - \frac{\vec{A}_{l,r}^{T} \cdot \vec{B}}{|\vec{B}|^2} \cdot \vec{B}, \qquad \vec{A} = \frac{1}{2} \cdot \left(\frac{\vec{A}_{l_p}}{|\vec{A}_{l_p}|} + \frac{\vec{A}_{r_p}}{|\vec{A}_{r_p}|} \right), \qquad \vec{H} = \frac{\vec{B}}{|\vec{B}|}, \qquad \vec{V} = \vec{A} \times \vec{H} \quad (1)$$

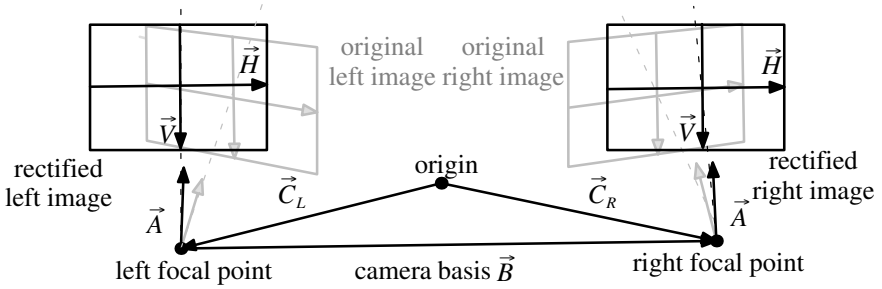


Fig. 2 Binocular image rectification

The images are rectified by projectively mapping all pels of the virtual image plane into the original image plane and interpolating the color information using a lowpass filter. Due to the fixed focal point, this projection is independent of the depth of the scene. An example of image rectification is shown in Fig 3.

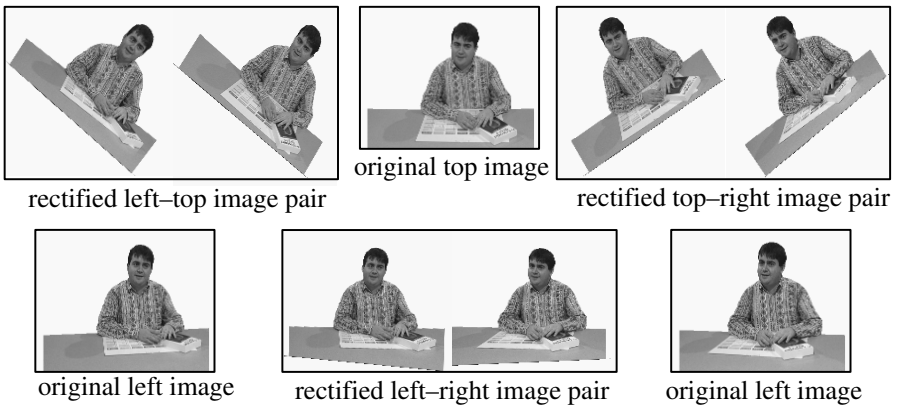


Fig. 3 Original image triple and the three rectified image pairs

BINOCULAR DEPTH ESTIMATION

For each rectified image pair, a disparity map is estimated using a block–matching technique that additionally considers relationships between neighbouring disparity values and estimates hierarchically on a multi–resolutional image pyramid [5], [6]. This approach expects rectified images and estimates all disparity values of a scan line simultaneously in order to consider the neighbourhood constraints. In addition to the

disparity values, it provides the disparity error variance σ_{de}^2 as local reliability term for each estimate. σ_{de}^2 is derived from the remaining luminance differences ΔI between two corresponding blocks and the luminance gradients inside the blocks $\delta I/\delta x$ (2), [7].

$$\sigma_{de}^2 = \frac{\sigma_{\Delta I}^2}{\sum \left(\frac{\delta I}{\delta x}\right)_i^2}, with \quad \sigma_{\Delta I}^2 = \text{remaining luminance difference variance} \tag{2}$$

Disparity is only estimated in high textured regions, where the camera noise variance is significantly lower than the signal variance along the epipolar line. This increases the robustness of the estimator, because the similarity measure is less affected by camera noise in high textured regions. Areas with small luminance variance are marked undetermined. The reliability of the estimates is further increased using a disparity cross–check between corresponding disparity values in a pair of disparity maps that is estimated of each rectified image pair using (3).

$$|d_L(x, y) - d_R(x - d_L, y)| \begin{cases} \leq 1 \Rightarrow d = estimated\ disparity \\ > 1 \Rightarrow d = undetermined \end{cases} \tag{3}$$

Depth values are computed from the estimated disparities by means of triangulation. In addition, σ_{de}^2 is converted into a depth error variance $\sigma_{\tau e}^2$ by projecting the standard deviation of the disparity error onto the line of sight of the image point. The resulting depth range on the line of sight models the standard deviation of the depth error and is squared to compute $\sigma_{\tau e}^2$.

BACK–PROJECTION AND AVERAGING OF DEPTH ESTIMATES

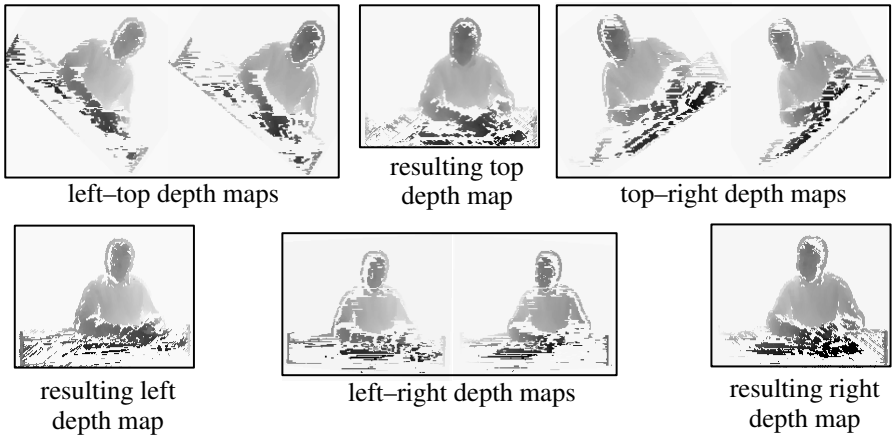


Fig. 4 Estimated depth maps of the three rectified image pairs and the resulting three back projected and averaged depth maps
■ =1.3m ■ =1.5m □ =1.7m □ =undetermined

For each image triple six depth maps are calculated, two for each of the rectified image pairs (Fig. 4). The estimated depth maps of the rectified image pairs and the

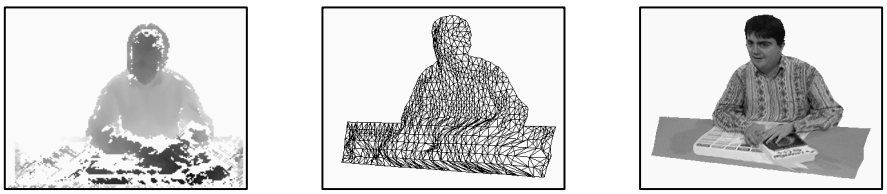
depth error variances are back–projected into the original image planes. This results in two depth maps and depth error variance maps for each original image. A resulting depth map τ_{av} is computed by averaging the two depth estimates τ_1 and τ_2 for each pel according to their reliabilities $\sigma^2_{\tau e_1}$ and $\sigma^2_{\tau e_2}$ using a Kalman filter (4).

$$\tau_{av} = \frac{\tau_1 \cdot \sigma^2_{\tau e_2} + \tau_2 \cdot \sigma^2_{\tau e_1}}{\sigma^2_{\tau e_1} + \sigma^2_{\tau e_2}} \quad , \quad \sigma_{\tau_{av}} = \frac{\sigma^2_{\tau e_1} \cdot \sigma^2_{\tau e_2}}{\sigma^2_{\tau e_1} + \sigma^2_{\tau e_2}} \tag{4}$$

RESULTS

The presented approach has been tested using synthetic noisy images of a single virtual object and real image sequences of video communication scenes. The synthetic images enable a numeric evaluation of the depth estimation error. Compared to the binocular approach, the trinocular approach reduces the depth error variance by 40%, due to averaging of two independently estimated depth maps for each image. This reduction is achieved independent of the adjusted amplitude resolution of the disparity estimates and the noise variance. E.g. the depth error variance is decreased from 0.44mm^2 down to 0.27mm^2 in case of quarter–pel resolution of the disparity estimates, a noise variance of 3, a distance between the virtual cameras of 0.6m and a distance between virtual object and virtual camera system of 1.7m. When averaging two statistically independent estimates with identical expected values of the estimation error, a reduction by 50% is expected. This reduction has not been achieved, because the averaged depth estimates and their estimation errors are not statistically independent. For example, the estimation error is partly caused by differences between model assumptions of the estimator and properties of the real objects, which are similar for both estimates.

The size of undetermined regions, resulting from an insufficient luminance structure or occlusions, can be reduced significantly. This reduction is visible in Fig. 4, when the density of the rectified depth maps is compared to the density of the resulting depth maps. There, the number of reliably estimated depth values is increased from approximately 50000 pel to 70000 pel. The increased density and accuracy improves the quality of the resulting 3–D model of the scene. An example is shown in Fig. 5.



Estimated depth map | shape of 3–D model from new point of view | textured 3–D model from new point of view

Fig. 5 Scene representation using textured 3–D model that is computed from the estimated depth maps

CONCLUSIONS

A block–based depth estimation approach has been presented that evaluates image triples captured with an unrestricted calibrated camera setup. From each image triple, three image pairs are formed and evaluated independently after respective rectification. Efficient binocular block–based depth estimation is used to compute a pair of depth maps for each rectified image pair. After back–projected into the original image planes this results in two estimated depth maps per image. These are averaged according to their depth error variances.

A comparison to binocular depth estimation using synthetic noisy images and real images showed that the trinocular approach provides more accurate and reliable depth estimates. In case of synthetic images, the depth error variance is reduced by 40% due to averaging the two independently estimated depth maps for each original image. For real video communication sequences, the number of reliably estimated depth values is increased by 40% due to the reduction of occlusions and due to the independence of luminance structure directions.

This work is supported by the EU in the European ACTS–PANORAMA project.

[1] O. Faugeras, "Three–Dimensional Computer Vision: A Geometrical Viewpoint", The MIT Press, Cambridge, Massachusetts, 1993, ISBN 0–262–06158–9.

[2] R. Y. Tsai, "A Versatile Camera Calibration Technique for High–Accuracy 3D Machine Vision Metrology Using Off–the–Shelf TV Cameras and Lenses", Journal of Robotics and Automation, Vol. RA–3. No.4, August 1987, pp. 323–344.

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

[4] R. Koch, "Automatische Oberflächenmodellierung starrer dreidimensionaler Objekte aus stereoskopischen Rundumansichten", Ph. D. Thesis, Uni Hannover, Hannover, Germany, 1996.

[5] L. Falkenhagen, "Depth Estimation from Stereoscopic Image Pairs Assuming Piecewise Continuos Surfaces", published in "Image Processing for Broadcast and Video Production", Springer Verlag, Workshops in Computing, Hamburg 1994, ISBN 3–540–19947–0.

[6] L. Falkenhagen, "Hierarchical Block–Based Disparity Estimation Considering Neighbourhood Constraints", submitted to Int. Workshop on Synthetic–Natural Hybrid Coding and Three Dimensional Imaging – IWSNHC3DI'97, Rhodes, Greece, 1997.

[7] D.E. Wells, "The method of least squares", Department of Surveying Engineering, University of New Brunswick, Fredericton, N.B. 1971.