SYSTEMS FOR DISPARITY-BASED MULTIPLE-VIEW INTERPOLATION

Jens-Rainer Ohm, Ebroul Izquierdo and Karsten Müller

Heinrich-Hertz-Institut, Image Processing Department
Einsteinufer 37, D-10587 Berlin, Germany

ABSTRACT

Viewpoint adaptation from multiple-viewpoint video captures is an important tool for telepresence illusion in stereoscopic presentation of natural scenes, and for the integration of real-world video objects into virtual 3D worlds. This paper describes different low-complexity approaches for generation of virtual-viewpoint camera signals, which are based on disparity-processing techniques and can hence be implemented with much lower complexity than full 3D analysis of natural objects or scenes. A realtime hardware system, which is based on one of our algorithms, has already been developed.

1. INTRODUCTION

Generation of changed view directions from video scenes or video objects is one central problem in interactive multimedia applications, e.g. when video data are composed with graphics material and the user is allowed to "navigate" or "fly" through a scene, or when several people cooperate in a virtual space [1]. If a multiview capture of a scene or an object is taken, the task of viewpoint adaptation can be accomplished by extracting information from the available camera views. The synthesis of natural-looking intermediate views can be done by interpolation from the different-view images, if the positions of corresponding points are known. This requires the knowledge of disparity data, which can immediately be used to project pixels onto an intermediate image plane. Section 2 describes an approach for low-cost disparity estimation ; the viability of this scheme has been shown by a hardware realization [2].

A critical case is the presence of occlusion areas, where some parts of the scene will only be found in one of the left- and right-view images. In these cases, disparities cannot be estimated, and instead of interpolation, a unidirectional projection has to be performed. This problem can be solved by application of foreground/background segmentation. Section 3 describes a concept for disparity-controlled scene segmentation.

We have investigated different techniques for viewpoint adaptation towards segmented video objects or foreground/background scenes, which are described in sections 4 and 5. The first is a scheme for disparity-controlled interpolation from stereoscopic image views [3]. The second is a technique, which first performs combination of the texture information from different views, and then projects this texture, also controlled by disparity data, to some specific viewpoint [4].

In section 6, conclusions are drawn.

2. DISPARITY ANALYSIS

If a video scene is acquired from two or more camera views, the appearance in the particular views is not identical. The shift between corresponding points in two views is called the *disparity*. If the geometry of the camera setup is known, it is possible to determine the absolute depth of a point in the scene. For example, when the optical axes of two cameras are parallel, the absolute depth in the scene is proportional to the reciprocal value of the disparity ; large disparities indicate that an object is very near to the cameras. For a more detailed treatise of this topic, the reader is referred to [5].

Disparity estimation is basically a task of correspondence matching, and is the most demanding task of the systems we describe. The disparity range to be used during estimation should be adequate to the interval between possible minimum and maximum disparities within the scene. We have found that we need disparity ranges of up to 120 pixel with a 50 cm baseline, and up to 230 pixel with an 80 cm baseline between cameras, when the minimum distance between cameras and an object was approximately 1.5 m.

During the last years, many different schemes for disparity estimation have been proposed. Though feature-based [6,7,8] and dynamic-programming [9] approaches seem to perform very well, we found them to be too complex for a hardware system with the requirement of large disparity ranges. Matching approaches can be classified as area-based schemes [10,11]. We have developed a hierarchical block matching scheme, which easily copes with arbitrary disparity ranges, and performs very robustly even in the case of low correspondence between left- and right-view images. A criterion based on an absolute-difference feature is used to determine optimum positions of the matching windows.

Our disparity estimation algorithm was developed under the constraint of hardware feasibility and was described in more detail in [2,3]. The system can be divided into 4 different modules :

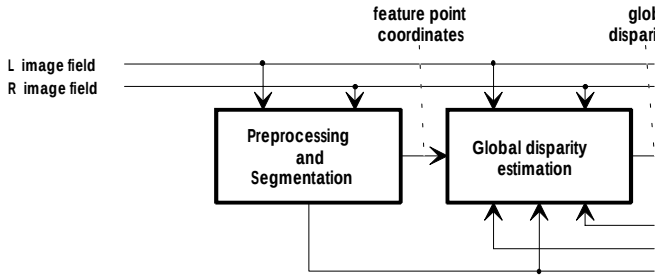

Fig.1. Block diagram of hierarchical block matching disparity estimator

1. Preprocessing and segmentation. The goal of this stage is to find points with highest relevance for matching, and to perform a subdivision into foreground and background areas.
2. Block matching with large block size for global bidirectional disparity estimation with cross-consistency check.
3. Block matching with small block size for local bidirectional disparity estimation with cross-consistency check.
4. Interpolation of dense L⟶R and R⟶L disparity fields, application of vertical median filters and ordering-constraint check.

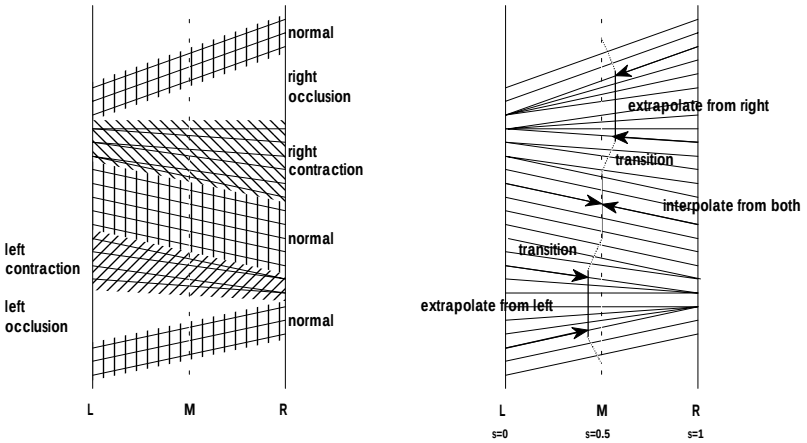

Fig.3. a definition of "normal", occlusion and contraction areas
b selective projection/interpolation within these areas

A flowchart describing the interrelation of the disparity estimator module blocks is given in fig.1. Preprocessing and segmentation is performed on both input signals. Bidirectional (L⟶R and R ⟶L) sparse disparity fields are estimated in the global and refined in the local estimation stage. In order to guarantee temporal continuity of the estimated disparities and to avoid temporally annoying artefacts, the disparities estimated for the previous field pair are fed back to the estimator stages. For this purpose, the dense field, generated at the final stage by bilinear interpolation, is used. The system has been realized in dedicated hardware for realtime processing of CCIR 601/656 TV resolution video, and is ready to work [2].

3. SCENE SEGMENTATION

Disparity analysis allows, as a byproduct, a very reliable separation of foreground objects from a scene, because usually there is a remarkable discontinuity in the disparity in the vicinity of the object border. In this case, a foreground/background separation can be performed even in the case when the foreground object is not moving. Fig. 2 shows a result of automatic foreground classification based on combined texture/disparity analysis. For exact contour position determination based on texture component, the morphological watershed was extracted [12].

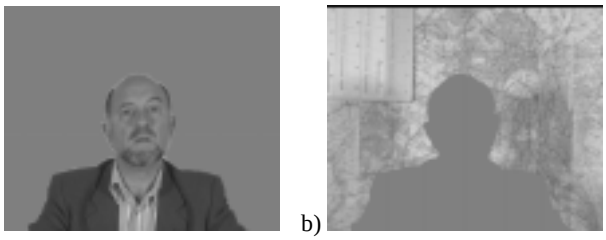

Fig.2. Result of combined disparity/texture segmentation
a foreground **b** background

4. STEREOSCOPIC INTERMEDIATE VIEW INTERPOLATION

The dense disparity fields provided by the estimator are used to project the left- and right-view images onto an arbitrary intermediate image plane. Herein, it has to be decided which areas of the intermediate image must be interpolated, i.e. taken from *both* images, and which areas are subject to occlusions and hence must only be projected from the corresponding area of *one* of the left- or right-view images. For both purposes, enhanced results can be obtained if the information from the segmentation mask is used. If we work with videoconferencing sequences, we can employ a very simple model for head-and-shoulder scenes, which is based on the more or less convex surface of the human head and body [13]. This model, however, also applies to a wide range of natural objects.

The relationship between corresponding points in the left and right image views is defined by the disparities. We can identify different cases, examples of which are given in fig. 3a. In a "normal" area, the disparity is constant, which indicates that the area is equally visible in both cameras. In a right or left occlusion case, the area is invisible in the left or right camera

view, respectively. In the case of a right or left contraction, the visibility resolution of the area from the respective camera is much higher, even though it also remains visible from the other camera. For a convex object, it is clear that left contractions/occlusions can occur only left from the object's center, while right contractions/occlusions will be present only right from this point. For example, the left-hand side of a person's head can be found with more accuracy in the left-view image, and vice versa, the right-hand side is better visible in the right-view image. In the center of the object, and also for the case of far background, we have "normal" areas, which means that texture information is equally visible in both cameras. Hence, for generation of the intermediate view, we proceed as indicated in fig.3b, where the information in occlusion and contraction areas is projected only from the left or right image view, while normal areas are interpolated from both views.

Fig. 4 shows - side by side - the original left camera view, an interpolated view in between both camera positions, and the original right camera view. This technique can also be applied separately to segmented foreground and background of a scene, which are then composed together into one output scene.

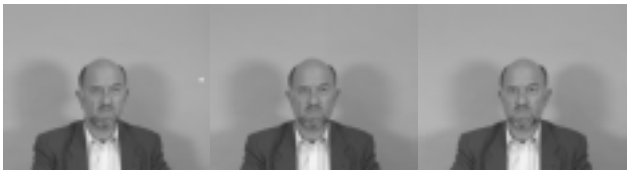

Fig.4. Result of object-based intermediate image interpolation (mid) compared to original left- and right-camera views

5. "INCOMPLETE 3D" TECHNIQUE

We have also developed a new technique denominated as *Incomplete 3D* (I3D) representation for video objects, which tries to combine the advantage of simplicity in disparity-based intermediate viewpoint interpolation with an approach for texture compression from multiple camera views. For that purpose, the texture surface of an object visible from two or more camera views is extracted and combined together, such that those points available in several views are only contained once, with the highest possible resolution. Basically, the technique to extract the texture surface from multiple camera views is also based on the "visibility" criteria described in the previous section. For the special case of a convex object (e.g. a head-and-shoulders scene), each camera has a unique area where it has the highest-resolution view towards the object. Hence, we define an "area of interest" (AOI) for each camera, and then simply "glue together" the AOIs of all cameras, taking into respect the disparity at the AOI border in the latter step.

The depth information is represented by a disparity map associated with the video object, such that it is not necessary to perform analysis of absolute depth or camera parameters. Viewpoint synthesis is achieved by disparity-controlled projection from the texture surface, separately within each AOI. Hence, disparity data must be available as part of the representation. Fig.5 shows the original left and right image views along with the extracted texture surface. Remark that the "incomplete 3D" surface is broader, because it bears more information than one of the single views. Fig.6 shows different view angles projected from the texture surface in fig.5. While the left and right images in the top line of fig.6 correspond to a synthesis of the original left and right view positions, the two images in the bottom line of the figure are synthesized positions beyond the original left and right camera views.

Fig.5. Left camera view, extracted "incomplete 3D" object and right camera view.

Fig.6. Different views synthesized from "incomplete 3D" object. Bottom pictures are from positions beyond the cameras.

We have performed experiments within the context of the MPEG-4 video verification model (VM) [14], where the surface texture data, as well as the outer (2D) shape of the unwrapped object's surface, are encoded as a video object plane (VOP). Disparity data representing the 3D shape are compressed by the mechanism included in the VM for encoding of the so-called graylevel alpha shape. This way, our incomplete 3D approach is fully compatible to the MPEG-4 video encoding technique (which was originally developed exclusively for 2D VOP data), retaining the 3D functionality of viewpoint adaptation. The lowest data rate we achieved was 56 kbit/s for the whole information, including disparity map, at CIF resolution of the sequence. The rate for the disparities is usually by a factor of 10 smaller than the rate necessary to encode the texture data.

The examples of fig.7 show MPEG-4 decoded results obtained with the I3D technique. Both I3D texture surface and disparity data were compressed with the MPEG-4 video VM software. These results indicate that the quality of viewpoint synthesis projection is independent of the strength of compression, which merely influences the sharpness of the image.

a)  b) 

Fig.7. "Incomplete 3D" video object decoded from MPEG-4 bitstream.
a Low compression (QP=1) **b** high compression (QP=16)

6. CONCLUSIONS

Disparity estimation is a key element in stereoscopic and multiple-camera analysis of 3D structures. If a processing of single objects within a scene is required, disparity data can be used to support segmentation, because they bear information about the 3D shape. The techniques and results presented in this paper show, that disparity-based techniques can directly be applied to accomplish the task of viewpoint adaptation towards single video objects or segmented foreground/background scenes. These approaches are much less complex than "genuine" 3D techniques like wireframe or 3D mesh modeling of video objects. The algorithm for disparity estimation is available as realtime hardware for TV resolution video. Hence, a realtime implementation of the techniques introduced in this paper is feasible.

ACKNOWLEDGEMENTS

This work has been supported by the German Federal ministry of education, research, science and technology under grant 01BN 701. Some stereo and multiview sequences were provided by CCETT/CNET, France.

REFERENCES

[1] "Immersive Telepresence," special issue of *IEEE MultiMedia Mag.*, vol. 4, no.1, Jan.-Mar. 1997

[2] J.-R. Ohm et al. : "A realtime hardware system for stereoscopic videoconferencing with viewpoint adaptation," *Image Communication*, special issue on 3D TV, January 1998

[3] J.-R. Ohm, E. Izquierdo : "An object-based system for stereoscopic viewpoint synthesis," *IEEE Trans. Circ. Syst. Video Tech.*, special issue on Multimedia Technology, vol. CSVT-7, no.5, pp. 801-811, Oct.1997

[4] J.-R. Ohm and K. Müller : "Incomplete 3D representation of video objects for multiview applications," *Proc. Picture Coding Symposium (PCS'97)*, pp. 427-432, Berlin, Sept. 1997

[5] O.D. Faugeras : "Three-Dimensional Computer Vision," MIT Press, Cambridge, Mass. : 1993

[6] W. Hoff and N. Ahuja : "Surfaces from stereo : Integrating feature matching, disparity estimation and contour detection," *IEEE Trans. Patt Anal. Mach. Intell.*, vol. PAMI-11, no.2, 1989.

[7] J. Weng, N. Ahuja and T.S. Huang : "Matching two perspective views," *IEEE Trans. Patt Anal. Mach. Intell.*, vol. PAMI-14, no.8, 1992.

[8] H.H. Baker and T.O. Binford : "Depth from edges and intensity based stereo," *Proc. 7th Int. Joint Conf. Artif. Intell.*, pp. 631-636, Vancouver, Canada, Aug. 1981.

[9] Y. Ohta and T. Kanade : "Stereo by intra- and inter-scanline," *IEEE Trans. Patt Anal. Mach. Intell.*, vol. PAMI-7, no.2, pp. 139-154, Mar. 1985.

[10] I.J. Cox, S.L. Hingorani and S.B. Rao : "A maximum likelihood stereo algorithm," *Computer Vision and Image Understanding* 63 (1996), no.3, pp.542-567

[11] B. Chupeau : "A multiscale approach to the joint computation of motion and disparity : Application to the synthesis of intermediate views," *Proc. 4th Europ. Worksh. on Three-Dimension. Televis.*, pp. 223-230, Rome, Italy, Oct. 1993.

[12] E. Izquierdo and S. Kruse : "Disparity-controlled segmentation," *Proc. Picture Coding Symposium (PCS'97)*, pp. 737-742, Berlin, Sept. 1997

[13] E. Izquierdo and M. Ernst : "Motion/disparity analysis for 3D video conference applications," *Proc. Int. Workshop on Stereoscopic and Three Dimensional Imaging*, pp. 180-186, Santorini, Greece, Sept. 1995.

[14] ISO/IEC JTC1/SC29/WG11 : "MPEG-4 video verification model version 8.0," Document no. N1796, July 1997