## Contact Details of Presenting Author

Edward Cooke (cooke@hhi.de)
Tel: +49-30-31002 613
Fax: +49-30-3927200

## Summation

- Examination of the representation of time-critical, arbitrary-shaped, video objects in 3-dimensional(3D) scenes.
- Discussion of problems with current 3D modeling techniques.
- Brief explanation of the recently developed *Incomplete 3D* technique.
- Outline of a demonstrator we developed which extends Incomplete 3D to present a new realtime, view adaptable reconstruction of a 3D object.
- Examination of the difficulties which arose during this work.
- Details of planned future work .

# Realtime View Adaptation of Video Objects in 3-Dimensional Virtual Environments

**Edward Cooke   Michael Karl   Peter Kauff   Oliver Schreer,**
**Heinrich-Hertz-Institute(HHI), Image Processing Department,**
**Alt-Moabit 74, D-10555 Berlin, Germany.**

## Abstract

In this paper we examine the representation of time-critical, arbitrary-shaped, video objects in 3-dimensional(3D) scenes. We outline the details of a demonstrator we developed in order to present a new method of generating a realtime, view adaptable reconstruction of a 3D object taken from 2 cameras. To date the modeling of a 3D object has either been too computationally expensive a process to allow for realtime; or been implemented as a simple texture mapping on a 2-dimensional node, which produces inconsistencies in the view adaptation during navigation in the scene. By extending the recently developed *Incomplete 3D* technique, a disparity-based multiview representation for a weakly convergent camera setup, we present an application which produces a realtime 3D consistent view from points outside the original camera baseline. An explanation of the difficulties which arose during this extension, and details of planned future work to extend the application to one containing a strongly convergent camera system are given.

## Introduction

We are currently seeing a huge growth in the amount of applications integrating arbitrary-shaped natural video objects into virtual environments(VE). This growth is due to the promotion of new international standards such as MPEG-4, and VRML, and the rapid growth of e-commerce, entertainment, and multimedia systems willing to exploit them. One of the main intentions of such applications and standards is to provide this integration of video objects seamlessly; and in the case of time-critical applications, in realtime. An example of such an application is the *Virtual Meeting Point* video conference application, Figure 1, which was developed at HHI in close co-operation with T-Nova Deutsche Telekom Innovationsgesellschaft [1].

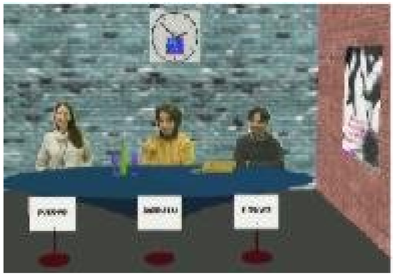

**Figure1:** Video communication in Virtual Meeting Point

## Inadequacies of Current Representations

In applications such as *Virtual Meeting Point* the viewpoint presentation quality of video objects during navigation is very important. The flat 2D nature of video objects may cause unrealistic geometric perspective views during navigation through the virtual 3D worlds. The solution is trivial when the 2D video object represents a scene element that is also 2D in reality.

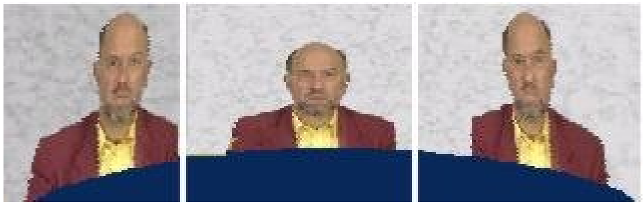

**Figure 2:** Left, centre, and right view of a 2D representation of a 3D object.

However, if the video object represents an element which is 3D in reality and acts as a 3D object in the virtual scene, eg a conference participant, perspective anomalies will occur, Figure 2. These anomalies can be avoided by using a Billboard node [2]. Billboard nodes are implemented in such a way as to automatically rotate to face the viewer in the VE, Figure 4. One advantage of this feature is that the total number of polygons in a world can be drastically reduced as geometric faces are texture mapped onto a 2D image and simply rotated around a particular axis. However in applications which require several video object representations of 3D objects and a realistic form of viewpoint adaptation, such as *Virtual Meeting Point*, the billboard node alone is not sufficient. One possibility to overcome this problem of viewpoint adaptation is to reconstruct the 3D shape of the objects from multi-viewpoint video signals. Two commonly used approaches are:

- *3D Modeling* - 3D models consist of deformable surface meshes or wireframes consisting of a set of adjacent elementary planar patches and are commonly used to describe surfaces with a desired precision. This can be achieved in principle as long as multiple camera views are available and the respective camera positions are known [ 3,4]. However, three drawbacks are associated with 3D models created from 2D views from different perspectives:
    - often incapable of properly reflecting the physical surface characteristics of the object, because the nodes and edges of the mesh generally have no physical significance;
    - a large number of small triangles are required at areas of high local curvature;
    - large algorithmic complexity.

In time critical applications, the complexity of the analysis of the video signals combined with generating the 3D models often overloads the realtime capability of the application.

- *Intermediate Viewpoint Interpolation* [ 5,6,7] - In this approach disparities are estimated from adjacent camera views, and an intermediate view is generated by disparity-compensated interpolation from the original views. There are two main problems with this approach:
    - the considerable increase in the bitrate to be coded and transmitted because one texture map per view, and disparity maps for the correspondences between these views, are required;
    - the disparity data derived for optimum encoding is often not appropriate for viewpoint interpolation [

      8].

## The Incomplete 3D Approach

We have decided to examine a hybrid solution, *Incomplete 3D* (IC3D). IC3D is a disparity-based multiview representation that was developed here at HHI in the context of MPEG-4. This *incomplete*ness is two-fold:

- the technique does not retain the full pixel representation of all the views available, thus resulting in higher compression;
- it does not perform full 3D modeling analysis, with the advantage of a simplified complexity.

The general concept is to limit the number of pixels that have to be encoded, by analysis of the correspondences between the particular views available, such that for an object, each area that is visible within more than one camera view is encoded only once with the highest possible resolution. Since the disparities indicate the correspondences of pixels between two camera views, it is straightforward that missing pixels of one view can be reconstructed by disparity compensated projection from the other view. Hence, the disparity data can also be used to reconstruct other views, e.g. by performing interpolation and/or extrapolation of the pixels from the available views.

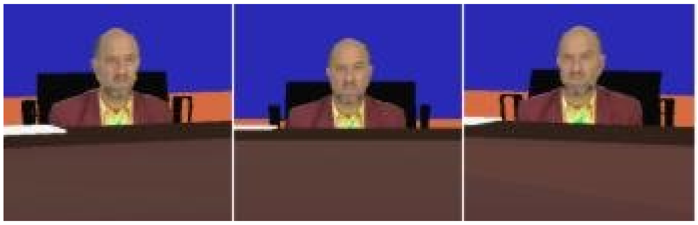

**Figure 3:** View adaptive synthesis using IC3D and a billboard node.

The IC3D model is displayed in the VE using a billboard node. The advantages of the billboard over a normal flat 2D object were discussed in *Inadequacies of Current Representations*. The rotation axis is set around the*Y* axis of the billboard node at the point of convergence, Figure 4. This combination of IC3D model with billboard node provides a more realistic viewpoint synthesis as the newly synthesised view is always rotated to face the viewer, Figure 3. The IC3D process is described in detail in [9].

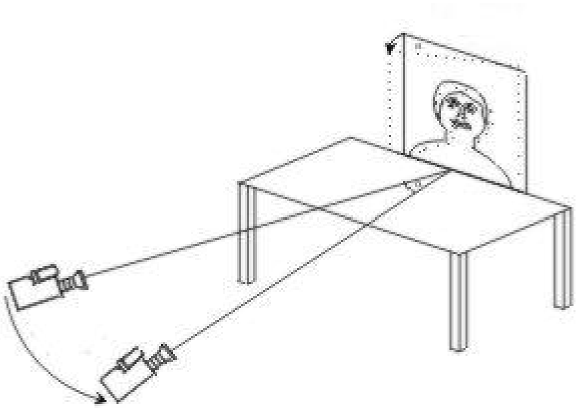

**Figure 4:** As the virtual viewpoint moves by angle $\alpha$ the billboard rotates and the corresponding image is synthesised

## Experimental Results

In this section we describe the modifications we made to the initial algorithm to develop a realtime, 3D viewpoint adaptive application using IC3D, and discuss the problems that arose during these modifications. The main challenge was the integration of the algorithm into a *live* application; previously it processed stored images which were captured using a parallel or near parallel camera setup. We developed a test application which uses a software based graphics renderer, has 2 cameras to capture video signals, and runs without any dedicated hardware. We achieved a video frame rate of 10Hz at QCIF resolution. Table 1 shows details of a complexity measure for all processes running on a Pentium 400 MHz dual processor machine.

| **Proceessing Step** | 3D Compositor | Segmentation | IC3D Analysis | IC3D Synthesis |
|---|---|---|---|---|
| **Time(ms)** | ~60 | 18 | 54 | 12 |

**Table 1:** Complexity measure.

These values indicate that the IC3D analysis step, depth analysis based on disparity estimation, and the 3D composition of the scene, clearly dominate the computing time of the demonstrator. It is clear that with dedicated graphics hardware and code optimized for a parallel architecture a frame rate of greater than 12.5Hz, and eventually CIF resolution, is attainable. In our demonstrator we use 2 cameras, and in order to capture enough object information to make viewpoint adaptation worthwhile, they have a relatively large baseline. The problem with a large baseline and the original parallel camera setup is that the amount of matching information contained in both left and right images is not optimal if the object is relatively small. This has implications for the position of the conference participant; as in our application the user had to sit well away from the cameras and hence the terminal in order to be in both left and right images, causing problems with respect to user keyboard and mouse access. In order to alleviate this problem we chose to use a convergent camera setup, which captures more information because the cameras are centered on or around the object. However, our new choice of setup affects the analysis and synthesis modules of the original IC3D algorithm. Using a parallel setup the adaptive viewpoint selection functions only along the baseline between the cameras. Once beyond this the quality of the synthesised image decreases rapidly due to lack of information, effectively restricting navigation in the 3D VE. In a convergent camera setup the geometric relationship between the local coordinate systems of two cameras is no longer described by a simple translation along the baseline. We must now also take into account the convergence angle, $\alpha$, of the 2 cameras, which gives us the possibility of moving outside the baseline, hence allowing much freer navigation in the VE. Due to restrictions of the initial algorithm that have not yet been overcome, and to allow us to make a direct comparison to the parallel optical axises set-up, we assume a weakly convergent system in our initial application. In both cases, vertical disparities are always equal to zero. Moreover, depth can be directly computed from horizontal disparity. However, it is important to notice that these similarities are of a formal nature only. In fact, the physical interpretations of these two systems are quite different:

- Firstly, horizontal disparities depend on the baseline in the case of parallel optical axises whereas they depend on an angle, $\alpha$, in convergent systems;
- Secondly, and more importantly, with parallel optical axises the horizontal disparity is inversely proportional to depth. Hence, zero-disparities only exist for points at infinity. While in a convergent system the opposite is true.

These differences are significant and provide us with a new approach for the synthesis of the viewpoint. In effect the convergent setup provides a way of computing a virtual view at any arbitrary position around the point of convergence. This allows us to navigate around the VE with the same freedom given by any other 3D model but with the advantage of much faster model generation.

## Future Work

So far, only the horizontal disparity shift has been considered in the disparity analysis. While this suffices for our weakly convergent setups it is not sufficient to solve the correspondence problem of strongly convergent setups. In this case the epipolar lines must be derived in the second image plane for each point in the first image plane or vice versa. This can be achieved if the orientations and positions of the cameras are known. In the case of a parallel setup, all epipolar lines are parallel, and it is only necessary to search for a corresponding point within the same scan line of the other image. Since the correspondence estimation is often performed by a matching operation, the reduction of necessary matches by introduction of the so-called epipolar constraint is important; at the same time, the accuracy of the estimated disparity field is increased.

## Conclusions

In this paper we discussed the current problems of a realtime 3D representation of video objects. IC3D was presented as a realtime 3D solution. The examples and results presented in this paper show how the integrity of the 3D model is preserved during user navigation at positions outside the original baseline. An application was developed to confirm the results for a weakly convergent camera setup. An explanation of the problems that occurred during this development, along with a description of how we intend to produce a demonstrator which functions for a strongly convergent setup were given. A presentation of the current demonstrator is planned to be given at the conference in accompaniment to this paper.

## Acknowledgements

This study is supported by the Ministry of Science and Technology of the Federal Republic of Germany, Grant-No.01BN701/1.

## Bibliography

[1] Stefan Rauthenberg et al. Virtual shop and virtual meeting point - two prototype applications of interactive services using the new multimedia coding standard mpeg-4. *ICCC*, 1999.

[2] VRML-Consortium.*ISO/IEC 14772-1:1997, The Virtual Reality Modeling Language*. The VRML Consortium, 1997.

[3] E. Izquierdo and X. Feng. *Image-Based 3D Modeling of Arbitrarily Natural Objects*. Urbana, Illinois, USA, October 1998. VLBV Work-shop.

[4] E. Izquierdo and X. Feng. Modeling of arbitrarily objects based on geometric surface conformity. *IEEE Transactions on CSVT*, February 1999. Special Issue on SNHC.

[5] E. Chen and L. Williams. View interpolation for image synthesis. In *Proc. ACM SIGGRAPH'93*, pages

[6] T. Werner, R.D. Hersch, and V. Hlavác. Rendering real-world objects using view interpolation. In *Proc. IEEE*

279-288, 1993.

[7] J.-R. Ohm and E. Izquierdo. An object-based system for stereoscopic viewpoint synthesis. *IEEE Trans. Circ. Syst. Video Tech.*, CSVT-7(5):801-811, October 1997.

[9] J.-R. Ohm and K Müller. Incomplete 3d for multiview representation and synthesis of video objects. *ECMAST98*, pages 26-41, 1998.

*Int. Conf. Comp. Vision*, pages 957-962, Boston, 1995.

[8] B.L. Tseng and D. Anastassoiu. Multiviewpoint video coding with mpeg-2 compatibility. *IEEE Trans. Circ. Syst. Video Tech.*, CSVT-6(4):414-419, August 1996.