International Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging IWSNHC3DI'99, Santorini, Greece, September 1999.

3-D IMAGING AND COMPRESSION – SYNTHETIC HYBRID OR NATURAL FIT?

Bernd Girod, Peter Eisert, Marcus Magnor, Eckehard Steinbach, Thomas Wiegand

Telecommunications Laboratory, University of Erlangen-Nuremberg
Cauerstrasse 7, 91058 Erlangen, Germany
{girod|eisert|magnor|steinb|wiegand}@nt.e-technik.uni-erlangen.de

Invited Paper

ABSTRACT

This paper highlights recent advances in image compression aided by 3-D geometry information. As two examples, we present a model-aided video coder for efficient compression of head-and-shoulder scenes and a geometry-aided coder for 4-D light fields for image-based rendering. Both examples illustrate that an explicit representation of 3-D geometry is advantageous if many views of the same 3-D object or scene have to be encoded. Waveform-coding and 3-D model-based coding can be combined in a rate-distortion framework, such that the generality of waveform coding and the efficiency of 3-D models are available where needed.

1. INTRODUCTION

Source models play an important role in image and video coding. Knowledge that is available a priori and that can be represented appropriately need not be transmitted. Rate distortion theory allows us to calculate a lower bound for the average bitrate of any coder, if a maximum permissible average distortion may not be exceeded. Often, practical schemes perform close to their rate-distortion theoretical bounds. It may not be concluded, however, that this fundamentally prevents us from inventing even more efficient coding schemes. Rate distortion theoretical bounds are valid only for a given source model, and often these models are rather crude. A more sophisticated source model might result in a lower rate at a given distortion. Better source models are the key to more efficient image compression schemes.

The majority of images are the result of a camera pointing to a three-dimensional scene. The scene consists mostly of surfaces reflecting the illumination towards the camera according to well understood physical laws. *Three-dimensional models* throughout this paper are models capturing the three-dimensional spatial structure of a scene in front of the camera along with the optical and photometric laws that govern the image formation process. Given that 3-D models seem such a natural fit for image compression, their success for this application has been remarkably poor. Almost all practical compression schemes are based on random process models that ignore the 3-D nature of the world being imaged.

The attempt to explicitly recover 3-D structure for a still image and use this information for coding is not very promising. The projection of the 3-D scene onto the image plane is an enormous data reduction, and a 3-D reconstruction has to overcome many ambiguities. How, for example, would one encode a (flat) photograph in a 3-D scene? We may, however, benefit from an explicit 3-D model when encoding a large set of 2-D images, where each individual image represents essentially the same 3-D scene, but possibly from a different viewing angle and/or at a different point in time. For example, for a video sequence resulting from a camera moving through a static 3-D (Lambertian) environment, we would ideally transmit a texture-mapped 3-D model of the environment once, and then only update the 3-D motion parameters of the camera.

In this paper, we show two examples of how explicit 3-D models can improve image compression. The first example, presented in Section 2, is a classic: model-based compression of head-and-shoulder views for videotelephony. The second example (Section 3) is an area of recently increased interest: compression of light fields.

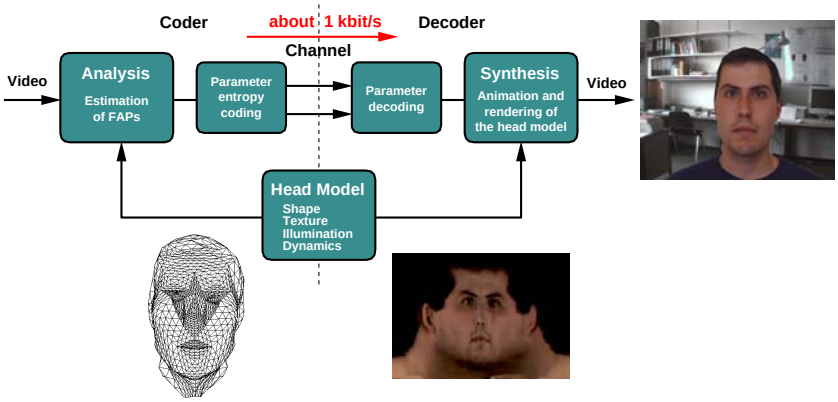


Figure 1: Basic structure of the model-based codec.

2. MODEL-AIDED COMPRESSION OF VIDEOPHONE SEQUENCES

For videotelephony, we want to transmit the head-and-shoulder view of a talking person. More than 15 years ago, Forchheimer et al. have proposed a videotelephone system based on a computer-animated 3-D head model [1] [2], and many groups have investigated such systems since [3]. Impressive progress has been made in the automatic tracking of facial expressions over the last few years [4]. For head-and-shoulder scenes, bit-rates of about 1 kbps with acceptable quality can be achieved. Unfortunately, a major drawback of such a system is still its limitation to a specific 3-D model and hence lack of generality.

In the following, we describe an extension of an H.263 video codec [5] that utilizes information from a model-based codec. Instead of exclusively predicting the current frame of the video sequence from the previously decoded frame, prediction from the synthetic frame of the model-based codec is additionally allowed. The encoder decides which prediction is more efficient in terms of rate-distortion performance. Hence, the coding efficiency does not decrease below H.263 in the case the model-based codec cannot describe the current scene. On the other hand, if the objects in the scene correspond to the 3-D models in the codec, a significant improvement in coding efficiency can be achieved.

2.1. Model-based Video Codec

The structure of a model-based codec is depicted in Fig. 1. The encoder analyzes the incoming frames and estimates the parameters of the 3-D motion and deformation of the head model. These deformations are represented by a set of facial animation parameters (FAPs) [6] that are entropy-encoded and transmitted through the channel. The 3-D head model and the facial expression synthesis are incorporated into the parameter estimation. The 3-D head model consists of shape, texture, and the description of facial expressions. For synthesis of facial expressions, the transmitted FAPs are used to deform the 3-D head model. Finally, individual video frames are approximated by simply rendering the 3-D head model.

In our model-based coder all FAPs are estimated simultaneously using a hierarchical optical flow based method starting with an image of 88 x 72 pixels and ending with CIF resolution. In the optimization an analysis-synthesis loop is employed [7]. The mean squared error between the rendered head model and the current video frame is minimized by estimating changes of the FAPs. To simplify the optimization in the high-dimensional parameter space, a linearized solution is directly computed using information from the optical flow and motion constraints from the head model. This approximative solution is used to compensate the differences between the video frame and the corresponding synthetic model frame. The remaining linearization errors are reduced by repeating the procedure at different levels of resolution. For more details about the model-based codec please refer to [4].

2.2. Proposed General Video Codec

Fig. 2 shows the architecture of the general, model-aided video codec (MAC). This figure depicts the well-known hybrid video coding loop that is extended by a model-based codec. The model-based codec is running simultaneously to

the hybrid video codec, generating a synthetic model frame. This model frame is employed as a second reference frame for block-based motion compensated prediction (MCP) in addition to the previously reconstructed reference frame. For each block the video coder decides which of the two frames to use for MCP. The bit-rate reduction for the proposed scheme arises from those parts in the image that are well approximated by the model frame. For these blocks, the bit-rate required for transmission of the motion vector and DCT coefficients for the residual coding is often highly reduced. For more details about the architecture and the mode-decision, see [8].

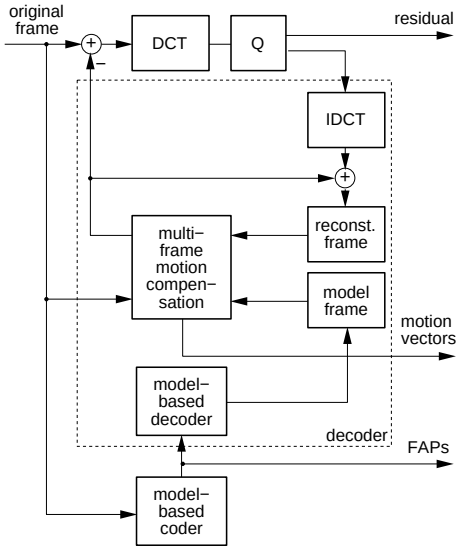


Figure 2: Structure of the proposed "model-aided" video coder. Traditional block-based MCP from the previous decoded frame is extended by prediction from the current model frame.

2.3. Experimental Results

Experiments are conducted for the standard CIF video test sequence *Akiyo*. The first 200 frames of this sequence are encoded at 10 Hz using both the H.263 and the model-aided H.263 coder. Since no head shape information from a 3-D scan is available for this sequence, a generic 3-D head model is used. Texture from the first video frame is mapped onto the object.

For comparison of the proposed coder with the anchor, the state-of-the-art test model of the H.263 standard (TMN-10), rate-distortion curves are generated by varying the DCT quantizer over the values $10, 15, 20, 25$, and 31. Bit-streams are generated that are decodable producing the same PSNR values as at the encoder. In our simulations, the data for the first INTRA frame and the initial 3-D model are excluded from the results. This way we simulate steady-state behavior, i.e., we compare the inter-frame coding performance of both codecs excluding the transition phase at the beginning of the sequence.

We first show rate-distortion curves for the proposed coder in comparison to the H.263 test model. The following abbreviations are used for the two codecs:

- **TMN-10:** The result produced by the H.263 test model, TMN-10, using Annexes D, F, I, J, and T.
- **MAC:** Model-aided H.263 coder: H.263 extended by model-based prediction with Annexes D, F, I, J, and T enabled as well.

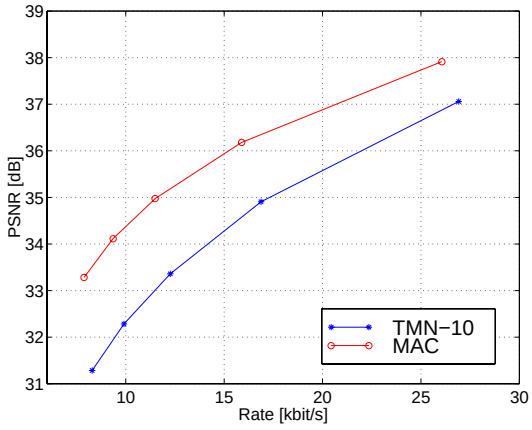


Figure 3: Rate-distortion plot for the video sequence *Akiyo*.

Fig. 3 shows the results obtained for the test sequence *Akiyo*. Significant gains in coding efficiency are achieved compared to TMN-10. Bit-rate savings of about 35 % at equal average PSNR are achieved at the low bit-rate end.

The upper half of Fig. 4 shows frame 150 of the TMN-10 coder, while the lower half corresponds to the model-aided coder. Both frames require about 720 bits. Significant visual improvements can be observed for the MAC codec. More experimental results can be found in [8].

3. LIGHT FIELD COMPRESSION

Light Field Rendering (LFR) constitutes a novel approach to generating arbitrary 2-D images of

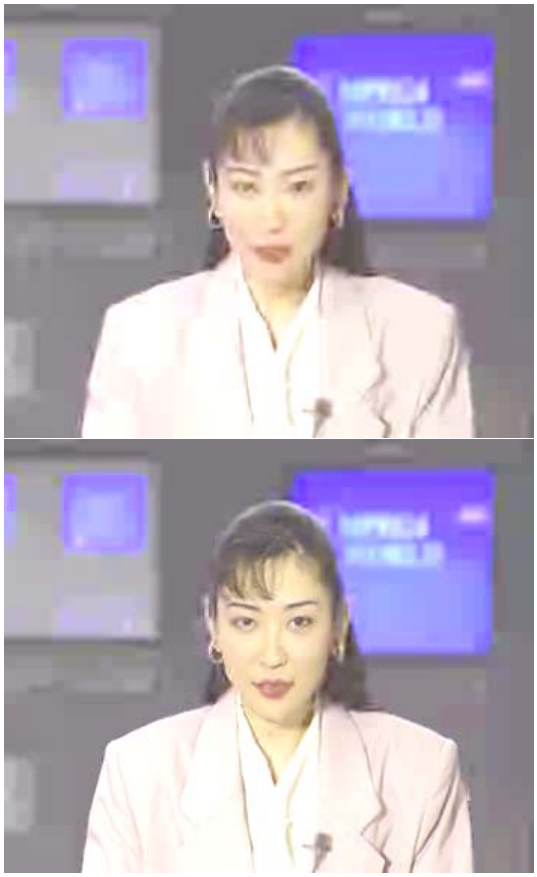


Figure 4: Frame 150 of the *Akiyo* sequence coded at the same bit-rate using the TMN-10 and the MAC, **upper image:** TMN-10 (31.08 dB PSNR, 720 bits), **lower image:** MAC (33.19 dB PSNR, 725 bits).

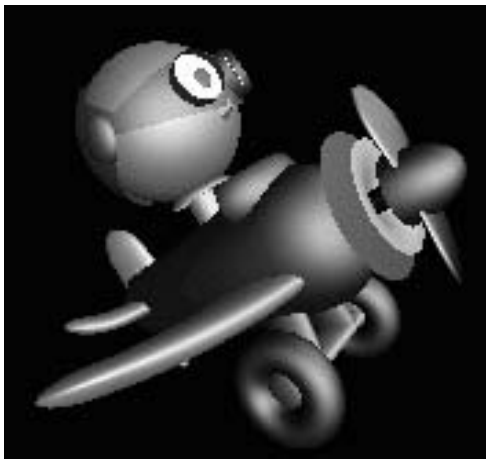


Figure 5: Image from the light field *Airplane*.

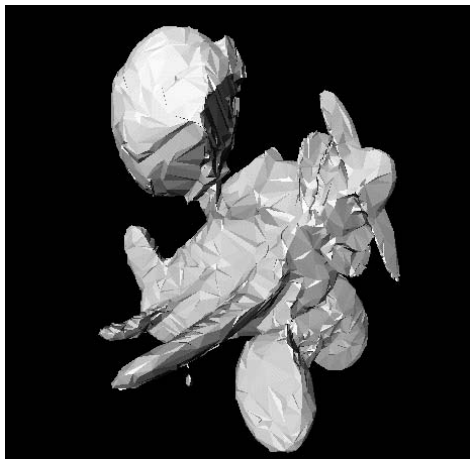


Figure 6: Reconstructed object geometry.

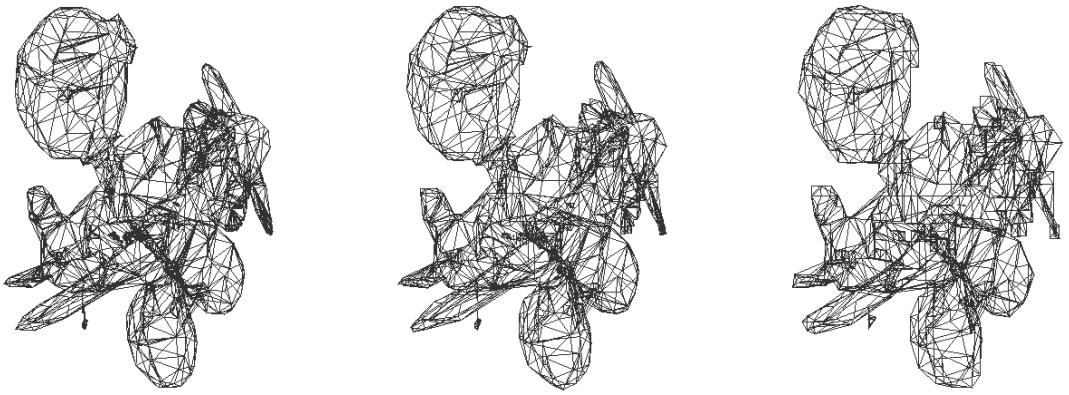
static 3-D scenes [9][10]. Traditional 3-D rendering relies on geometry models, textures and lighting descriptions. In LFR, the scene's visual appearance from multiple viewpoints, its *light field*, serves as basis for the rendering process. For a more detailed description of LFR please refer to [11] in this proceedings volume. A light field consists of a 2-D array of conventional 2-D images. Hence, light fields are a 4-D data set. To attain photorealistic rendering results, light fields must typically contain several thousand images, making data compression necessary for rendering, storing and transmitting light fields.

3.1. Image-based Compression

Because light fields consist of image data, still-image coding techniques are applicable to light fields. Vector quantization [9] and DCT-coding [12] have been employed to light-field coding, yielding compression ratios up to 30 : 1. Much higher compression ratios can be attained if inter-image similarities are considered. Compression techniques developed for video coding have been suitably modified for light-field compression in the block-based codec described in detail in [11] in this volume, achieving compression ratios up to 1000 : 1 with acceptable reconstruction quality.

3.2. Geometry-aided Compression

Light-field coding can benefit further from information about object geometry to compensate disparity between images. Because light fields do not contain explicit scene geometry, geometry information has to be inferred from the



5195 triangles
52594 bits
0.2% max. deviation

2582 triangles
27121 bits
0.8% max. deviation

1799 triangles
18516 bits
1.6% max. deviation

Figure 7: Wireframe model of the approximate object geometry, coded with the EMC algorithm at different resolution levels and bit-rates. The maximum deviation of vertex position is measured relative to the object's extension.

light-field images. Disparity maps can be derived for accurate motion compensation between neighboring light-field images [13], yet images farther away can only be motion-compensated at lower image resolution. If the scene exhibits texture or silhouette information, approximate 3-D object geometry can be reconstructed from the light field. The additional geometry information needs to be efficiently compressed to aid in light-field compression. A full 3-D geometry model offers the advantage of enabling disparity compensation of arbitrarily many light-field images over any distance at constant geometry coding bit-rate.

3.3. Experimental Results

The *Airplane* light field is used to show the advantages of using 3-D geometry in light field compression. Fig. 5 depicts one of the 8×8 light field images. Each image consists of 256×256 24-bit RGB pixels. The multi-hypothesis reconstruction algorithm described in [14] is used to build a volumetric model of the object. The volume surface is triangulated using a refined Marching Cubes algorithm, and the Progressive Meshes algorithm [15] is applied to simplify the triangle mesh without compromising model accuracy. The model shown in Fig. 6 is then coded using the Embedded Mesh Coding (EMC) algorithm described in [16]. The EMC algorithm allows efficient coding of arbitrary triangle meshes at variable resolution. Fig. 7 shows the approximate geometry at different resolution levels.

The geometry model is used to compensate disparity between light-field images. An efficient coding order is established by quadtree-decomposition of the light-field image array (Fig. 8): Only the array's 4 corner images are INTRA-coded using a standard block-based DCT scheme (images A in Fig. 8). The center image (image B in Fig. 8) is first disparity-compensated from the 4 corner images using the

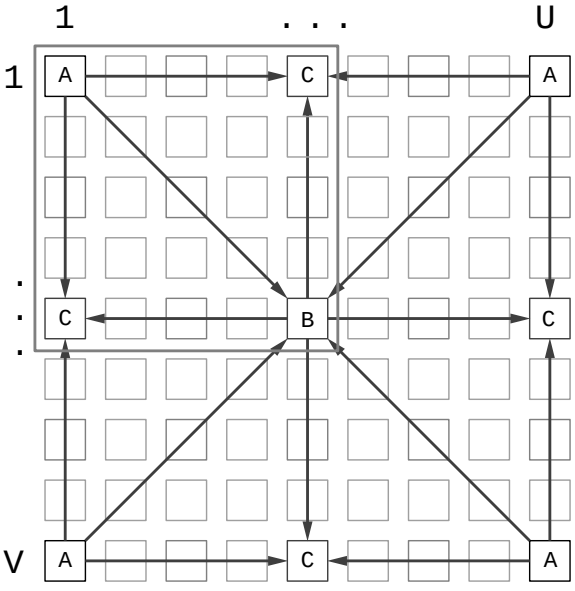


Figure 8: Disparity-compensation order of the light-field images: from the corner images (A), the center image (B) is predicted. The images at the middle of the sides (C) are compensated from the center image and the two closest corner images. The array is subdivided into quadrants and each quadrant is coded likewise. The algorithm keeps recursing until all images are coded.

geometry model, and the residual error is DCT-coded. The middle images on the array sides (images C in Fig. 8) are compensated from the center image and the two closest corner images, and the residual error is coded. The array is then divided into 4 quadrants, and in each quadrant the center image and the side images are coded as before. The algorithm steps recursively through the quadtree structure until all images are coded.

Fig. 9 shows the resulting distortion, measured as the average peak-signal-to-noise ratio (PSNR) over all light field images vs. bit-rate. For comparison, the block-based coder's RD-curve for the *Airplane* light field is depicted. The geometry coder yields $10-40\%$ better compression, depending on reconstruction quality. Even higher coding gains from approximate geometry can be expected for light fields consisting of more images, as can be seen from Fig. 9, if the additional bit-rate for coding the geometry information is neglected.

4. CONCLUSIONS

We have considered two very different applications in this paper: the model-based compression of head-and-shoulder video sequences and the compression of 4-D light fields. Both applications have in common that essentially the same 3-D object is visible in many 2-D images, from different viewing angles or at different time instances with deformation. We found that in both scenarios, an explicit geometry model helps to reduce the bit-rate. As the overhead for encoding the geometry information is distributed over a large number 2-D views, geometry-aided compression becomes increasingly attractive. An unresolved question is the minimum number of views, beyond which geometry-aided encoding is superior.

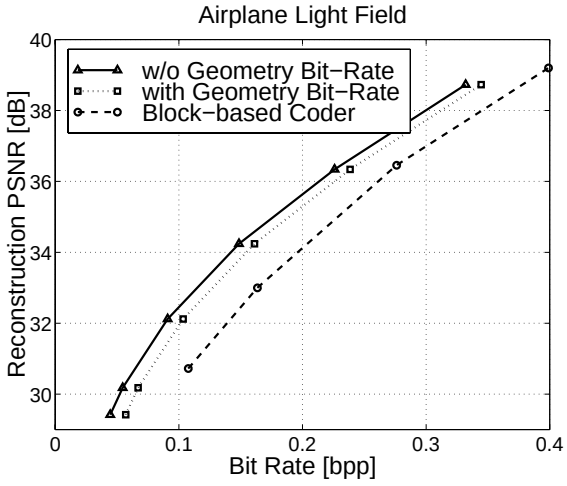


Figure 9: Rate-Distortion curves for the *Airplane* light field; the approximative geometry model is used to accurately compensate disparity, yielding better coding performance.

Our examples also illustrate that waveform-coding and 3-D model-based coding are not competing alternatives but should be combined to support and complement each other. Both can be elegantly combined in a rate-distortion framework, such that the generality of waveform coding and the efficiency of 3-D models are available where needed.

5. REFERENCES

[1] R. Forchheimer, O. Fahlander, and T. Kronander, "Low bit-rate coding through animation", *Proc. International Picture Coding Symposium PCS'83*, pp. 113–114, Mar. 1983.

[2] R. Forchheimer, O. Fahlander, and T. Kronander, "A semantic approach to the transmission of face images", *Proc. International Picture Coding Symposium PCS'84*, number 10.5, Jul. 1984.

[3] D. E. Pearson, "Developments in model-based video coding", *Proceedings of the IEEE*, vol. 83, no. 6, pp. 892–906, Jun. 1995.

[4] P. Eisert and B. Girod, "Analyzing facial expressions for virtual conferencing", *IEEE Computer Graphics and Applications*, vol. 18, no. 5, pp. 70–78, Sep. 1998.

[5] ITU-T Recommendation H.263 Version 2 (H.263+), "Video Coding for Low Bitrate Communication", Jan. 1998.

[6] *ISO/IEC FDIS 14496-2, Generic Coding of audio-visual objects: (MPEG-4 video), Final Draft International Standard, Document N2502*, 1999.

[7] H. Li, P. Roivainen, and R. Forchheimer, "3-D motion estimation in model-based facial image coding", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 545–555, Jun. 1993.

[8] P. Eisert, T. Wiegand, and B. Girod, "Rate-distortion-efficient video compression using a 3-D head model", *Proc. In-*

ternational Conference on Image Processing ICIP '99, Kobe, Japan, Oct. 1999.

[9] M. Levoy and P. Hanrahan, "Light field rendering", *SIGGRAPH 96 Conference Proceedings*, pp. 31–42, Aug. 1996.

[10] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph", *SIGGRAPH 96 Conference Proceedings*, pp. 43–54, Aug. 1996.

[11] M. Magnor and B. Girod, "Adaptive block-based light field coding", *Proc. International Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging IWSNHC3DI'99,* Santorini, Greece, Sept. 1999, this volume.

[12] G. Miller, S. Rubin, and D. Ponceleon, "Lazy decompression of surface light fields for precomputed global illumination", *Proc. Eurographics Rendering Workshop,* Vienna, Austria, pp. 281–292, Oct. 1998.

[13] M. Magnor and B. Girod, "Hierarchical coding of light fields with disparity maps", *Proc. International Conference on Image Processing ICIP-99,* Kobe, Japan, Oct. 1999.

[14] P. Eisert, E. Steinbach, and B. Girod, "Multi-hypothesis volumetric reconstruction of 3-D objects from multiple calibrated camera views", *Proc. International Conference on Acoustics, Speech, and Signal Processing ICASSP'99* Phoenix, USA, pp. 3509–3512, Mar. 1999.

[15] H. Hoppe, "Progressive meshes", *SIGGRAPH 96 Conference Proceedings*, pp. 99–108, Aug. 1996.

[16] M. Magnor and B. Girod, "Fully embedded coding of triangle meshes", *Proc. Vision, Modeling, and Visualization VMV'99,* Erlangen, Germany, Nov. 1999.