3-D IMAGE MODELS 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 bit-rate 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. An earlier version of this paper has appeared in [1]. Relative to [1], we have included new results in Section 3.

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 [2] [3], and many groups have investigated such systems since [4]. Impressive progress has been made in the automatic tracking of facial expressions over the last few years [5]. 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 [6] that utilizes information from a model-based codec. Instead of exclusively predicting the current frame of the video sequence from the previous decoded frame, prediction from the synthetic frame of the model-based codec is additionally allowed. The encoder decides which predic-

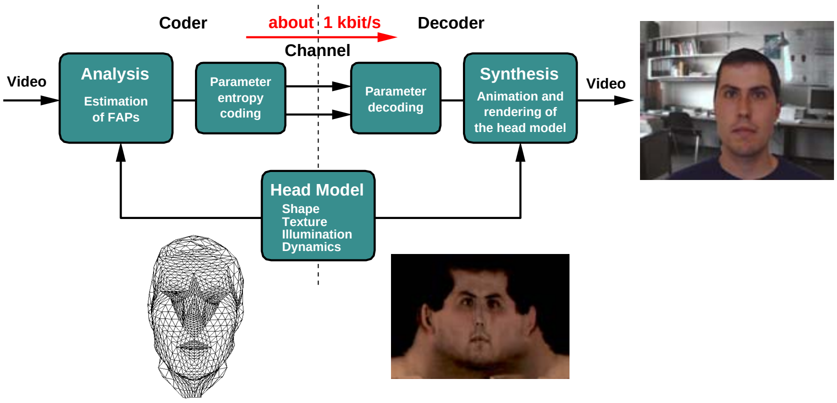


Figure 1: Basic structure of the model-based codec.

tion is more efficient in a rate-distortion sense. Hence, the coding efficiency does not decrease below H.263 when 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) [7] 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 [8]. 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 [5].

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 previous 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 [9] in these proceedings.

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 parameter 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-coded 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:

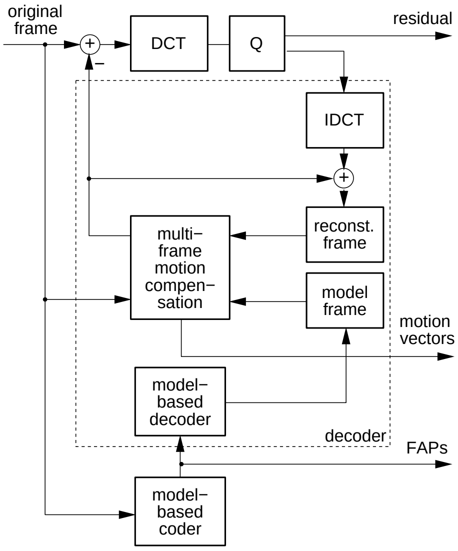


Figure 2: Structure of the proposed "model-aided" video coder. Traditional block-based MCP from the previous decoded frame is extended by prediction from the current model frame.

- **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.

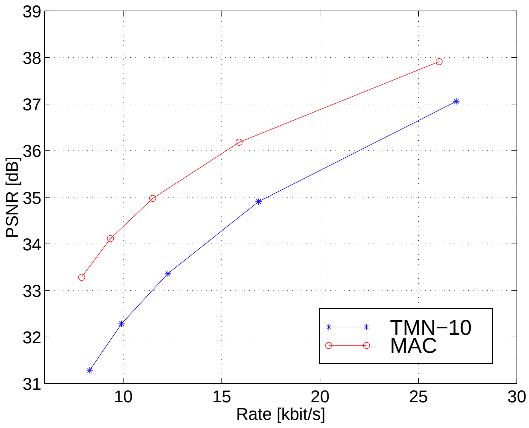


Figure 3: Rate-distortion plot for the video sequence *Akiyo*.

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 [9].

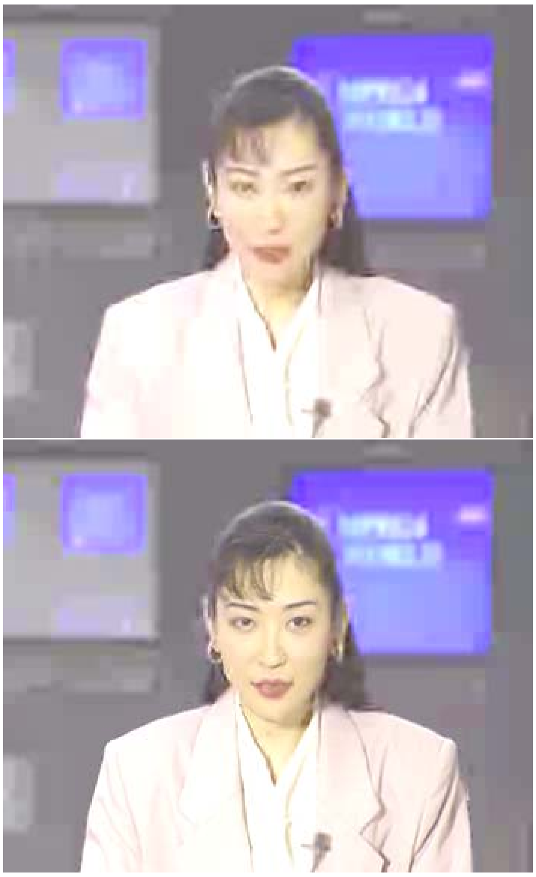


Figure 4: Frame 150 of the *Akiyo* sequence coded at the same bit-rate using the TMN-10 and the MAC, **upper image:** TMN-10 (31.08 dB PSNR, 720 bits), **lower image:** MAC (33.19 dB PSNR, 725 bits).

3. LIGHT FIELD COMPRESSION

Light Field Rendering (LFR) constitutes a novel approach to generating arbitrary 2-D images of static 3-D scenes [10, 11]. Traditional 3-D rendering relies on geometry models, textures and lighting descriptions. In LFR, the scene's visual appearance from multiple viewing directions, its *light field*, serves as basis for the rendering process. A light field consists of an array of conventional 2-D images. To attain photorealistic rendering results, light fields typically contain several thousand images, making data compression necessary for rendering, storing and transmitting light fields.

Even though LFR does not depend on object geometry, light-field coding can benefit from geometry information to compensate disparity between images. Geometry has to be inferred from the light-field images, as light fields do not contain explicit scene geometry. Disparity maps can be

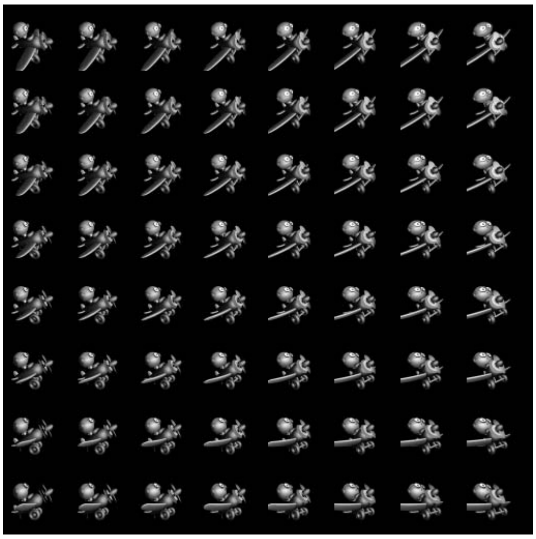


Figure 5: The *Airplane* light-field consists of a 2-D array of 8 × 8 images.

derived for accurate disparity compensation between neighboring light-field images (see [12] in these proceedings), yet images farther away can only be disparity-compensated at lower image resolution. If the light-field scene exhibits texture or silhouette information, approximate 3-D object geometry can be reconstructed. 3-D geometry aids in coding of light fields by enabling disparity compensation of arbitrarily many light-field images over any distance at constant geometry coding bit-rate.

Light-field coding with geometry information is demonstrated using the *Airplane* light field shown in Fig. 5. First, an approximate geometry model is derived from the 8 × 8 light-field images using the reconstruction algorithm described in [13] (Fig. 6). To code the approximate geometry model with adjustable accuracy, the Embedded Mesh Coding (EMC) algorithm described in [14] is employed to the reconstructed *Airplane* geometry. Fig. 7 shows the model coded at different resolutions. The geometry is used to generate high-resolution disparity maps for all images, and the light field is hierarchically coded as described in [12].

In Fig. 8, geometry-aided coding performance is compared with results from the disparity-map coder described in [12] and a block-based light-field coder [15]. Bit-rate for coding the geometry model is neglected to illustrate the possible coding gain for light fields consisting of several thousand images: full 3-D geometry yields up to 25% better compression over block disparity maps, and up to 40% lower bit-rate is achieved compared to block-based light-field coding. If coding the geometry model is taken into account, coding gain from explicit 3-D geometry compared to [12] is minor for the *Airplane* light field because of the low number of images to be coded. Fig. 9 depicts coding performance for different geometry resolutions if geometry bit-rate is included. Note that this geometry bit-rate also includes the "backside" of the airplane that is never visible in Fig. 5. Depending on reconstruction quality, optimum coding performance is achieved with different resolution models. While low-resolution geometry models do not compensate disparity well and might even introduce additional error, too detailed geometry degrades coding performance due to the increased geometry coding bit-rate. Optimum bit-rate allocation between geometry and residual-error coding is attained by selecting model resolution depending on overall target bit-rate.

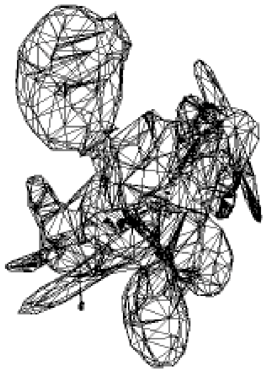
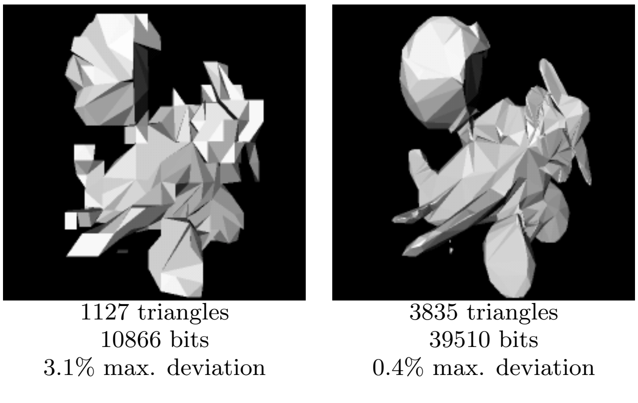


Figure 6: Approximate geometry model, reconstructed from the *Airplane* light-field images.

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 of geometry information is distributed over a large



1127 triangles
10866 bits
3.1% max. deviation

3835 triangles
39510 bits
0.4% max. deviation

Figure 7: *Airplane* geometry model, coded at different resolutions; maximum vertex displacement is measured relative to model size.

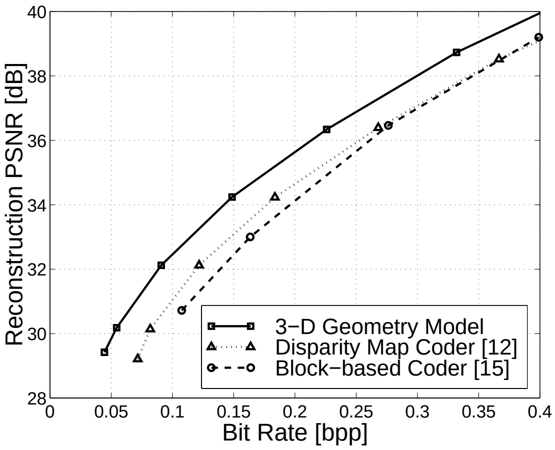


Figure 8: Rate-Distortion curves of different coders for the *Airplane* light field; geometry bit-rate is neglected.

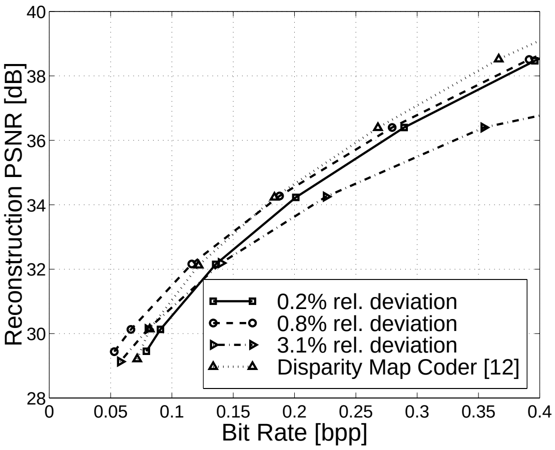


Figure 9: Rate-Distortion curves for different model resolutions if geometry bit-rate is taken into account.

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.

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] B. Girod, P. Eisert, M. Magnor, E. Steinbach, and T. Wiegand, "3-D imaging and compression - synthetic hybrid or natural fit?", *International Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging IWSNHC3DI'99, Santorini, Greece*, Sep. 1999.

[2] 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.

[3] 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.

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

[5] 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.

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

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

[8] 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.

[9] P. Eisert, T. Wiegand, and B. Girod, "Rate-distortion-efficient video compression using a 3-D head model", *Proc. International Conference on Image Processing ICIP '99,* Kobe, Japan, Oct. 1999.

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

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

[12] 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.

[13] 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.

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

[15] 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.