MODEL-AIDED CODING: USING 3-D SCENE MODELS IN MOTION-COMPENSATED VIDEO CODING

Peter Eisert, Thomas Wiegand

Telecommunications Laboratory
University of Erlangen-Nuremberg
{eisert,wiegand}@LNT.de

Bernd Girod

Information Systems Laboratory
Stanford University
girod@ee.stanford.edu

ABSTRACT

We show that traditional waveform-coding and 3-D model-based coding are not competing alternatives but should be combined to support and complement each other. Both approaches are combined such that the generality of waveform coding and the efficiency of 3-D model-based coding are available where needed. The combination is achieved by providing the block-based video coder with a second reference frame for prediction which is synthesized by the model-based coder. Since the coding gain of this approach is directly related to the quality of the synthetic frame, we have extended the model-aided coder [1] to exploit knowledge about illumination changes and multiple objects. Remaining model failures and objects that are not known at the decoder are handled by standard block-based motion-compensated prediction. A Lagrangian approach is employed to control the coder. Experimental results show that bit-rate savings of about 35 % are achieved at equal average PSNR when comparing the model-aided codec to TMN-10, the state-of-the-art test model of the H.263 standard.

1. INTRODUCTION

In recent years, several video coding standards such as H.261, H.263, MPEG-1, and MPEG-2 have been introduced, which mainly address the compression of generic video data for digital storage and communication services. These schemes are designed on the basis of the statistics of the video signal without knowledge of the semantic content and can therefore robustly be used for arbitrary scenes. The design of model-based codecs [2] is based on the semantics of the scene. Hence, if the semantic information of the scene can be exploited, higher coding efficiency may be achieved by model-based video codecs. For example, 3-D models that describe the shape and texture of the objects in the scene could be used. The 3-D object descriptions are encoded only once. When encoding a video sequence, individual video frames are characterized by 3-D motion and deformation parameters of these objects. In most cases, the parameters can be transmitted at extremely low bit-rates.

Such a 3-D model-based coder is restricted to scenes that can be composed of objects that are known by the decoder. One typical class of scenes are head-and-shoulder sequences which can be frequently found in applications such as video-telephone or video-conferencing systems. For head-and-shoulder scenes, bit-rates of about 1 kbit/s with acceptable quality can be achieved [3]. This has also motivated the recently determined *Synthetic and Natural Hybrid Coding* (SNHC) part of the MPEG-4 standard [4].

The combination of traditional hybrid video coding methods with model-based coding has been proposed by Chowdhury et al. in 1994 [5]. In [5] a *switched model-based coder* is introduced that decides between the encoded output frames from an H.261 block-based and a 3-D model-based coder. However, the mode decision is only done for a complete frame and therefore the information from the 3-D model cannot be exploited if parts of the frame cannot be described by the model-based coder. An extension to the switched model-based coder is the *layered coder* published by Musmann in 1995 [6]. The layered coder chooses the output from up to five different coders. The mode decision between the layers is also done frame-wise or object-wise and no combined encoding is performed.

In [1] we have presented an extension of an H.263 video coder [7] that utilizes information from a model-based coder. 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 coder is additionally allowed. The *model-aided coder* decides which prediction is efficient in terms of rate-distortion performance. Hence, the coding efficiency does not decrease below H.263 in the case the model-based coder cannot describe the current scene. On the other hand, if the objects in the scene are compliant to the 3-D models in the codec, a significant improvement in coding efficiency can be achieved.

In this paper we extend the model-aided coder [1] to exploit the efficiency of the model-based coder also for more sophisticated video sequences with changing lighting conditions or multiple objects. Parameters describing the illumination conditions in the scene are estimated together with motion and deformation of the objects resulting in more accurate model frames. Experimental results demonstrate that the improved rate-distortion performance of the model-aided codec remains also for these more general video sequences.

2. VIDEO CODING ARCHITECTURE

Figure 1 shows the architecture of the proposed model-aided video coder. This figure depicts the well-known hybrid video coding loop that is extended by a model-based coder. The model-based coder is running simultaneously to the hybrid coder, generating a synthetic model frame. This

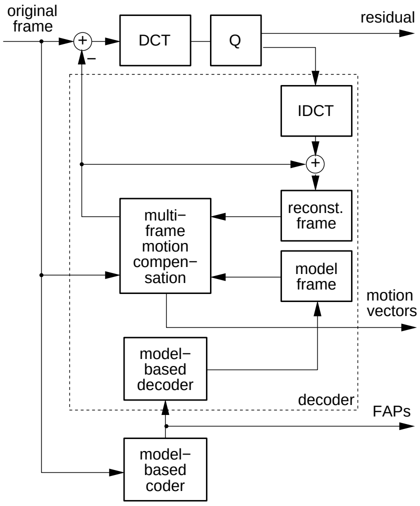

Figure 1: Structure of the model-aided video coder. Traditional block-based MCP from the previous decoded frame is extended by prediction from the current model frame.

model frame is employed as a second reference for block-based motion-compensated prediction (MCP) in addition to the previous reconstructed reference frame. For each macroblock, 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 rate-distortion optimized mode decision and the changes made to the H.263+ syntax, see [1].

3. MODEL-BASED CODEC

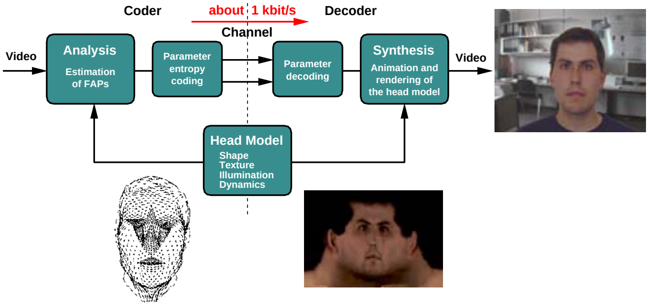

Figure 2: Basic structure of the model-based codec.

The structure of the model-based codec is depicted in Fig. 2. The encoder analyzes the incoming frames and estimates the parameters of 3-D motion and deformation for all objects in the scene. The deformations for the head model are represented by a set of facial animation parameters (FAPs) according to the MPEG-4 standard [4]. Motion and deformation of other objects are parameterized similarly. All parameters are then entropy-encoded and transmitted through the channel. The information from the 3-D models and the facial expression synthesis are incorporated into the parameter estimation. These models describe the shape, texture, and the motion constraints of the objects. For synthesis of facial expressions, the transmitted FAPs are used to deform the 3-D head model. The other objects are similarly moved and deformed in the virtual scene. Finally, individual video frames are approximated by simply rendering the 3-D scene.

In our model-based coder all parameters are estimated simultaneously using a hierarchical optical flow based method. In the optimization an analysis-synthesis loop is employed. The mean squared error between the rendered scene and the current video frame is minimized by estimating changes of the FAPs and the parameters for the other objects. 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 models. For the determination of those areas in the frame that correspond to a particular model, knowledge from the synthetic 3-D scene is exploited. The resulting 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 parameter estimation, please refer to [3].

In the case the lighting in the scene changes, the coder cannot represent the original video frame correctly. To overcome this restriction, we add an illumination component to the scene model that describes the photometric properties for colored light and surfaces. The incident light in the original scene is assumed to consist of ambient light and a directional light source with illumination direction $\mathbf{l}$. The surface is modeled by Lambertian reflection, and thus the relation between the video frame intensity I and the corresponding value I_{model} from the head model is

$$\begin{aligned} I^R &= I^R_{model}(c^R_{amb} + c^R_{dir} \cdot \max\{-\mathbf{n}\cdot\mathbf{l}, 0\}) \\ I^G &= I^G_{model}(c^G_{amb} + c^G_{dir} \cdot \max\{-\mathbf{n}\cdot\mathbf{l}, 0\}) \\ I^B &= I^B_{model}(c^B_{amb} + c^B_{dir} \cdot \max\{-\mathbf{n}\cdot\mathbf{l}, 0\}), \qquad (1) \end{aligned}$$

with c_{amb} and c_{dir} controlling the intensity of ambient and directional light, respectively [8]. The surface normal $\mathbf{n}$ is derived from the 3-D head model. The Lambertian model is applied to all three RGB color components separately with a common direction of the incident light. The equations (1) thus contain 8 parameters (ambient light: 3, directional light: 3, illumination direction: 2) that characterize the current illumination. By estimating these parameters using a linear least-squares estimator [8] we are able to compensate brightness differences of corresponding points in the synthesized model frame and the camera frame.

4. EXPERIMENTAL RESULTS

Experiments are conducted with the two self-recorded natural CIF sequences *Clapper Board* and *Illumination*. Rate-distortion curves are measured by varying the DCT quantizer parameter over values $10, 15, 20, 25,$ and 31. Bit-streams are generated that are decodable producing the same PSNR values as at the encoder. The data for the first INTRA frame and the initial 3-D model are excluded from the results thus simulating 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.

To specify the coding performance of the proposed model-aided codec, we compare it to the H.263 test model, TMN-10. The following abbreviations are used for the two cases:

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

For the special case of head-and-shoulder sequences, bit-rate savings of 35 % at the low bit-rate end corresponding to a coding gain of 2-3 dB PSNR are reported [1]. If the lighting in the scene changes this coding gain is reduced, since the model frames no longer describes the original video frames correctly. The additional estimation of the lighting situation, however, allows to adapt the illumination condition in the synthetic scene to the real world.

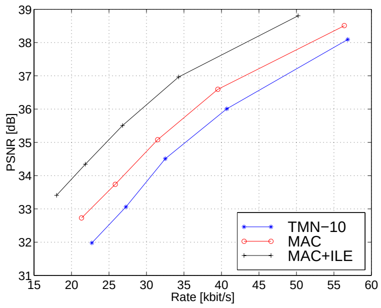

Figure 3: Rate-distortion plot for the sequence *Illumination* illustrating the achieved improvement when using an illumination estimator (ILE).

The effectiveness of the illumination estimation is illustrated in Fig. 3 for the sequence *Illumination*. During the acquisition of this sequence, one light source was moved to alter the illumination conditions. Two experiments are performed. For the first one, only the FAPs are estimated to create a model frame. For the second experiment, we additionally estimate the illumination parameters and generate motion- and illumination-compensated model frames. As shown in Fig. 3, the gain in PSNR for the model-aided coder compared to the TMN-10 is about 1 dB if no illumination compensation is performed. However, an additional gain of about 1.5 dB is achieved when exploiting illumination information.

In a second experiment, the influence of unknown objects in the scene is investigated. Fig. 4 shows the first frames of the head-and-shoulder sequence *Clapper Board*. During the first 50 frames, the face is occluded by an object that cannot be represented by the 3-D models available at the decoder. As a result, the corresponding model frames do not contain this additional object as shown in the upper image of Fig. 7. Since prediction from the previous decoded frame and residual coding provides us with robustness against model failures, the model-aided coder represents the entire frame correctly (second image in Fig. 7). The coding efficiency of the model-aided coder, however, drops down during the first frames as shown in the temporal evolution of the PSNR in Fig. 5 if the model-frame cannot represent the scene correctly. If the face is visible again the model frame can be exploited and the PSNR recovers showing high coding gains. The overall rate-distortion performance for the entire sequence is depicted in Fig. 6. Bit-rate savings of 33 % at the low bit-rate end are achieved.

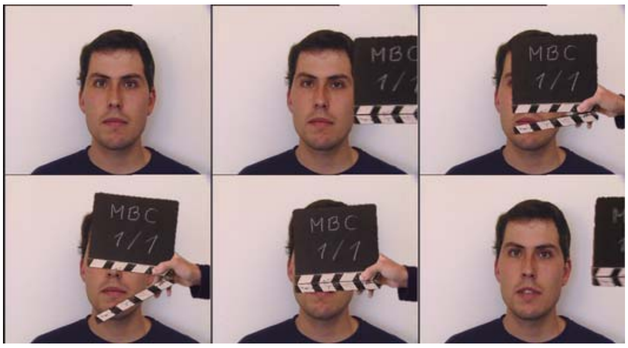

Figure 4: Frames 0, 11, 22, 33, 44, and 55 of the sequence *Clapper Board*.

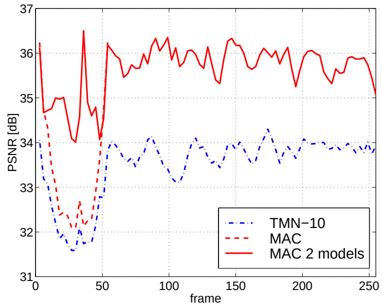

Figure 5: Temporal evolution of PSNR for the sequence *Clapper Board*. Both coders use a DCT quantizer parameter of 31.

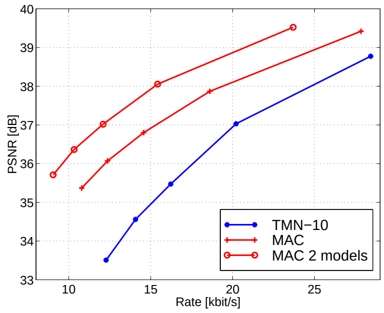

Figure 6: Rate-distortion plot for sequence *Clapper Board*.

Figure 7 illustrates the quality of the reconstructed frames for this case. The second image shows frame 54 encoded with the model-aided coder, while the next image corresponds to the TMN-10 coder at the same bit-rate. The lowest image of Fig. 7 shows a frame from the TMN-10 coder that has the same PSNR as the model-aided frame. Even though the PSNR is the same, the subjective quality of the reconstructed frame from the model-aided coder is

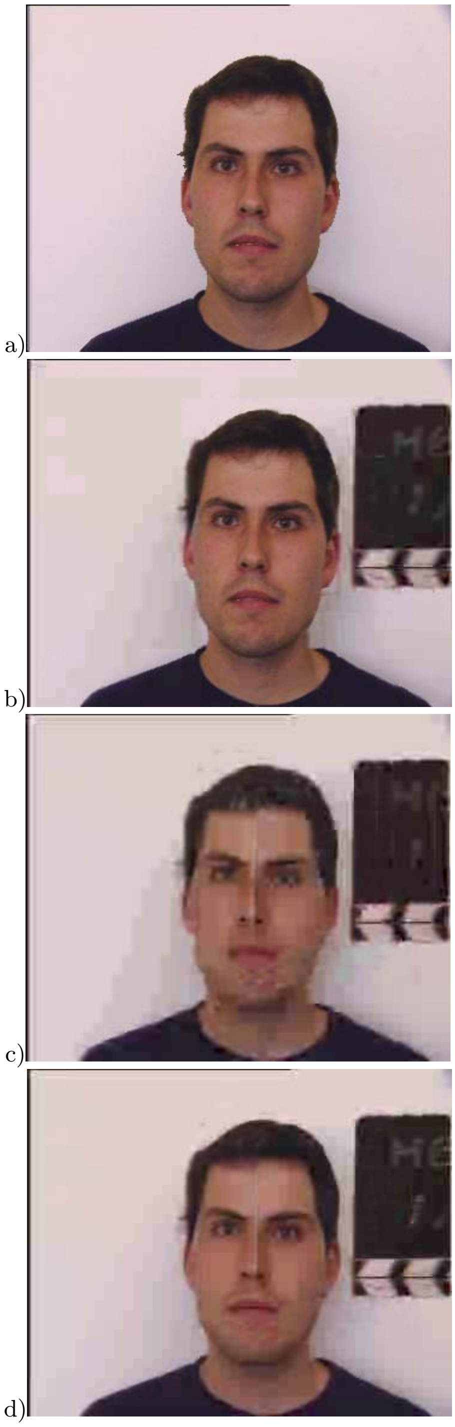

Figure 7: Frame 54 of the sequence *Clapper Board*. a) Model frame with missing clapper board; b) MAC, 36.1 dB, 2900 bits; c) TMN-10 with same bit-rate as MAC, 32.7 dB, 3200 bits; d) TMN-10 with same PSNR as MAC, 36.0 dB, 5800 bits.

clearly superior since facial features are reproduced more accurately and with less artifacts. The difference is even more striking when viewing motion sequences [1].

In a third experiment, the clap in the sequence *Clapper Board* is described by an additional 3-D model placed in the synthetic scene. This model is manually acquired using the texture from one frame that shows the entire clap. Motion and deformation parameters are estimated for both objects using the approach in Section 3. With these parameters, model frames are generated that represent all objects in the scene. Running the model aided coder with these model frames results in a much higher PSNR for the first frames compared to the case when using only the head model. This is illustrated in the upper curve of Fig. 5. The overall rate-distortion performance for the entire sequence is depicted in Fig. 6. Bit-rate savings of 45 % corresponding to a coding gain of about 3.5 dB are achieved.

5. CONCLUSIONS

The combination of model-based video coding with block-based motion-compensated prediction yields a superior video coding scheme for head-and-shoulder sequences. The advantages of both approaches are combined in a new framework by employing the synthesized frame from the model-based coder as a second reference frame for rate-constrained block-based motion-compensated prediction in addition to the previously reconstructed reference frame. Experiments have shown that the coding gain increases if more semantic information about the scene is available. This is exploited in the model-aided coder by estimating changes in the scene illumination and allowing multiple objects to move and deform. Bit-rate savings of about 35 % are achieved at equal average PSNR. When encoding at equal average bit-rate, significant improvements in terms of subjective quality are visible.

6. REFERENCES

[1] P. Eisert, T. Wiegand, and B. Girod, "Rate-distortion-efficient video compression using a 3-D head model", in *Proc. International Conference on Image Processing (ICIP)*, Kobe, Japan, Oct. 1999, vol. 4, pp. 217–221.

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

[3] 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.

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

[5] M. F. Chowdhury, A. F. Clark, A. C. Downton, E. Morimatsu, and D. E. Pearson, "A switched model-based coder for video signals", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 4, no. 3, pp. 216–227, June 1994.

[6] H. G. Musmann, "A layered coding system for very low bit rate video coding", *Signal Processing: Image Communication*, vol. 7, no. 4-6, pp. 267–278, Nov. 1995.

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

[8] P. Eisert and B. Girod, "Model-based coding of facial image sequences at varying illumination conditions", in *Proc. 10th Image and Multidimensional Digital Signal Processing Workshop IMDSP '98*, Alpbach, Austria, Jul. 1998, pp. 119–122.

[1] http://www.LNT.de/~eisert/mac.html