RATE-DISTORTION-EFFICIENT VIDEO COMPRESSION USING A 3-D HEAD MODEL

Peter Eisert, Thomas Wiegand and Bernd Girod

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

ABSTRACT

In this paper we combine model-based video synthesis with block-based motion-compensated prediction (MCP). Two frames are utilized for prediction where one frame is the previous decoded one and the other frame is provided by a model-based coder. The approach is integrated into an H.263-based video codec. Rate-distortion optimization is employed for the coding control. Hence, the coding efficiency does not decrease below H.263 even if the model-based coder cannot describe the current scene. On the other hand, if the objects in the scene correspond to the model-based coder, significant gains in coding efficiency can be obtained compared to TMN-10, the test model of the H.263 standard. This is verified by experiments with natural head-and-shoulder sequences. Bit-rate savings of about 35 % are achieved at equal average PSNR. When encoding at equal bit-rate, significant improvements in terms of subjective quality are visible.

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 utilize the statistics of the video signal without knowledge of the semantic content of the frames and can therefore robustly be used for arbitrary scenes.

In case the semantic information of the scene can be exploited, higher coding efficiency may be achieved by model-based video codecs [1, 2]. 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]. SNHC allows the transmission of a 3-D face model that can be animated to generate different facial expressions.

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. The decision which frame to take is based on rate and distortion. 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] as well as Kampmann and Ostermann in 1997 [7]. 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 this paper we present an extension of an H.263 video coder [8] 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 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.

This paper is organized as follows. We first describe the architecture of the video coder that combines the traditional hybrid video coding loop with a model-based coder that is able to encode head-and-shoulder scenes at very low bit-rates. The underlying semantic model is presented and the algorithm for the estimation of Facial Animation Parameters (FAPs) is briefly explained. We show how the information from the model-based coder is incorporated into the H.263 video coding scheme and how bit allocation is done in a rate-distortion-efficient way. Experimental results finally demonstrate the improved rate-distortion performance of the proposed scheme compared to TMN-10.

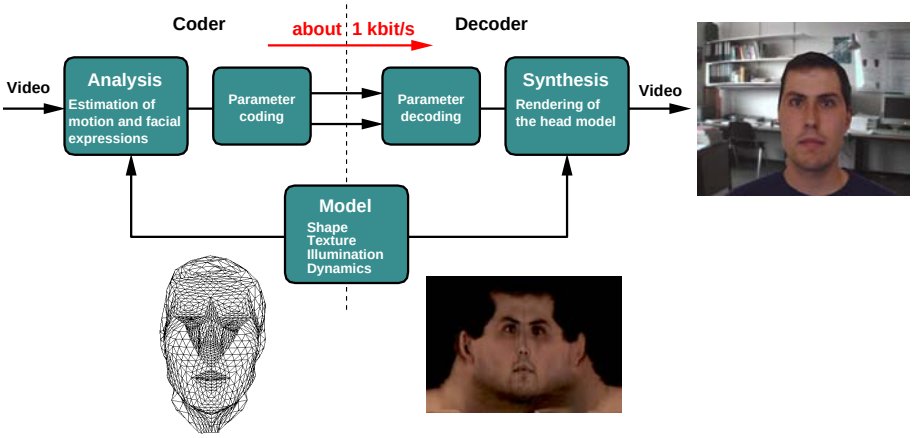

Figure 1: Basic structure of our model-based coder.

2. VIDEO CODING ARCHITECTURE

Figure 2 shows the proposed architecture of the hybrid model-based video coder. This figure depicts the well-known

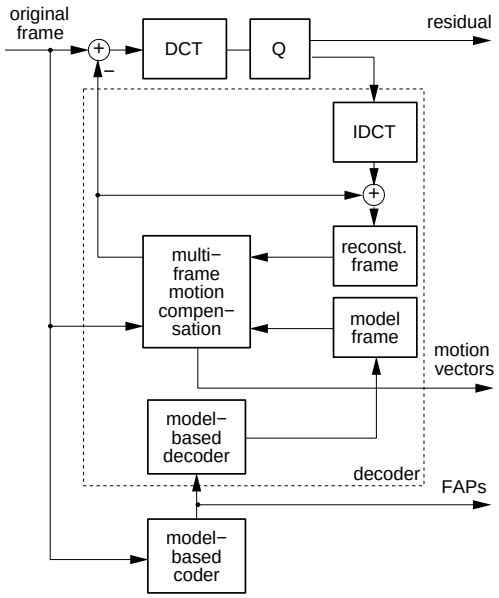

Figure 2: Structure of the video coder. Traditional block-based MCP from the previous decoded frame is extended by prediction from the current model frame.

hybrid video coding loop that is extended by a model-based coder. The model-based coder is running parallel to the hybrid coder, generating synthetic frames that are encoded at extremely low bit-rates. This synthetic approximation of the current frame is used as a second frame for the block-based motion-compensated prediction. For each macroblock the video coder decides which of the two frames is used. 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 transmission of the motion vector and the DCT-coefficients for the residual coding can often be avoided.

The H.263+ syntax is modified in that changes are made to the inter-prediction modes INTER, INTER-4V, and UNCODED[1] in order to enable multi-frame motion-compensated prediction. The INTER and UNCODED mode are assigned one code word representing the picture reference parameter for the entire macroblock. The INTER-4V mode utilizes four picture reference parameters each associated to one of the four 8×8 motion vectors.

3. MODEL-BASED CODER

The structure of our model-based coder is depicted in Fig. 1. The encoder analyzes the incoming frames and estimates the 3-D motion and deformation of all objects in the scene. Assuming that head and shoulder of a person are visible and can be described by our model, the estimator yields a useful set of facial expression parameters. In general, only a few parameters have to be encoded and transmitted, leading to very low bit-rates, typically less than 1 kbit/s [3]. At the decoder, the parameters are used to deform our head-and-shoulder model. This way the head-and-shoulder scene is approximated by simply rendering a 3-D model.

3.1. Head Model

For the description of the shape and color of head and shoulder, we use a generic model with fixed topology similar to the well-known Candide model [9]. In contrast to the Candide model, the object's surface is modeled with triangular B-splines [10] in order to reduce the number of degrees of freedom. This simplifies the modeling and estimation of facial expressions. For rendering purposes, the B-spline surface is finally approximated by a triangular mesh that is shown in Fig. 3. To initially adapt the generic model to the person we use a 3-D laser scan that provides us with information about the shape and color.

[1]The UNCODED mode is an INTER mode for which the COD bit indicates copying the macroblock from the previous frame without residual coding [8].

Figure 3: **left:** hidden-line representation of the head model, **right:** corresponding textured version.

For the estimation of the FAPs, we must be able to animate our model to create different facial expressions. This task is simplified by the use of B-splines because they already constrain the motion of neighboring vertices. For the parameterization of the facial expressions, the proposal of the MPEG-4/SNHC group [11] is used. According to that scheme, every facial expression can be generated by a superposition of 66 action units. These include both global motion, like head rotation, and local motion, like eye or mouth movement.

3.2. Facial Parameter Estimation

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 [12]. 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 the 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 coder please refer to [3].

3.3. Transmission of Facial Animation Parameters

In our experiments, 19 FAPs are estimated. These parameters include global head rotation and translation (6 parameters), movement of the eyebrows (4 parameters), two parameters for eye blinking, and 7 parameters for the motion of the mouth and the lips. For the transmission of the FAPs, we predict the current values from the previous frame, scale and quantize them. An arithmetic coder that is initialized with experimentally determined probabilities is then used to encode the quantized values. Note that the training set for the arithmetic coder is separate from the test set. The resulting bit-stream has to be transmitted as side information if the synthetic frame is employed for prediction.

4. RATE-CONSTRAINED CODER CONTROL

The problem of optimum bit allocation to the motion vectors and the residual coding in any hybrid video coder is a non-separable problem requiring a high amount of computation. To circumvent this joint optimization, we split the problem into two parts: motion estimation and mode decision.

The H.263-based multi-frame predictor conducts rate-constrained block-based motion estimation using both reference frames producing a motion-compensated frame. The motion estimation minimizes a Lagrangian cost function that is given by

$$J_{MOTION} = D_{DFD}(v) + \lambda_{MOTION} R(v), \qquad (1)$$

where the distortion D_{DFD} is measured as the sum of the absolute differences (SAD) and the rate term R is associated to the motion vector v.

Given the motion vectors, rate-distortion optimal macroblock modes are chosen. Rate-constrained mode decision minimizes

$$J_{MODE} = D_{REC} + \lambda_{MODE} R_{TOTAL}, \qquad (2)$$

for each macroblock [13]. Here, the distortion after reconstruction D_{REC} measured as the sum of the squared differences (SSD) is weighted against bit-rate using a Lagrange multiplier λ_{MODE}. The rate term is given by the total bit-rate R_{TOTAL} that is needed to transmit a particular macroblock mode, including the rates for the macroblock header, motion and texture coding.

The Lagrange multiplier for the mode decision is chosen following [14] as

$$\lambda_{MODE} = 0.85 Q^2. \qquad (3)$$

For the Lagrange multiplier used in the motion estimation, we make an adjustment to the relationship to allow use of the SAD measure. Experimentally, we have found that an effective such method is to measure distortion during motion estimation using SAD rather than the SSD and to simply adjust the Lagrange multiplier for the lack of the squaring operation in the error computation, as given by $\lambda_{MOTION} = \sqrt{\lambda_{MODE}}$.

5. EXPERIMENTAL RESULTS

Experiments are conducted on the self-recorded natural CIF sequence *Peter* and the standard video test sequence *Akiyo*. Both sequences consist of 200 frames and are encoded at 8.33 Hz and 10 Hz, respectively. 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. For the transmission of the face model description that is not changed during the sequence, the position of 316 control points and a texture map of size 450 x 512 pixels have to be encoded.

We first show rate-distortion curves for the proposed coder in comparison 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: H.263 extended by model-based prediction.

Figures 4 and 5 show the results obtained for the two sequences *Peter* and *Akiyo*. Bit-rate savings of about 35 % at equal average PSNR are achieved at the low bit-rate end.

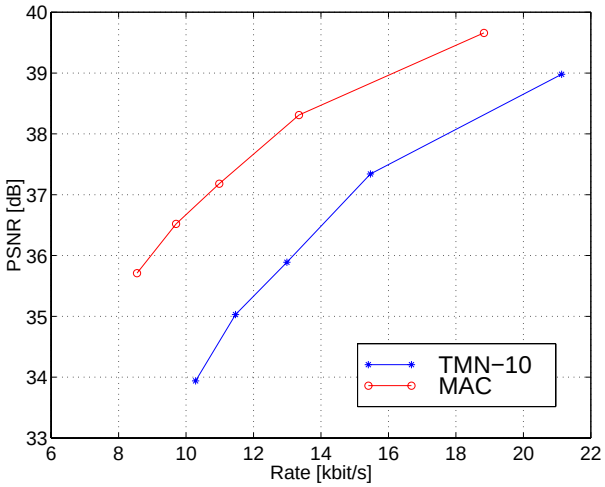

Figure 4: Rate-distortion plot for the sequence *Peter*.

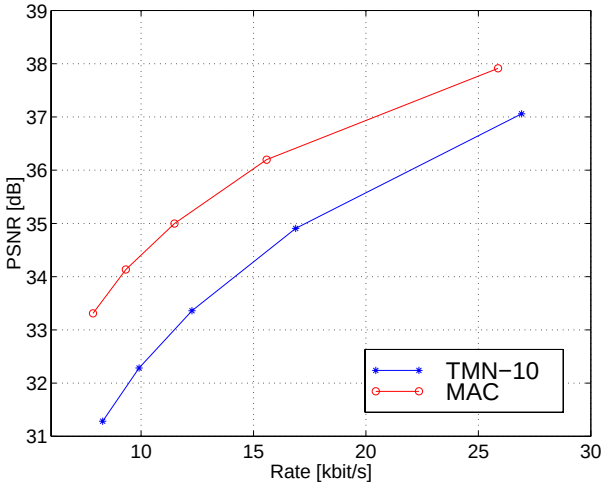

Figure 5: Rate-distortion plot for the sequence *Akiyo*.

Figure 6 shows frame 120 of the test sequence *Peter* where the upper picture is decoded and reconstructed from the TMN-10 decoder, while the lower one comes from the model-aided coder. Both frames are transmitted at the same bit-rate. The model-aided coder yields a PSNR improvement of about 3.5 dB.

In Figure 7 similar results for the sequence *Akiyo* are shown. The upper image corresponds to frame 150 that is decoded and reconstructed from the TMN-10 decoder, while the lower one is generated from the model-aided coder.

Figure 6: Frame 120 of the *Peter* sequence coded at the same bit-rate using the TMN-10 and the MAC, **upper image:** TMN-10 (33.88 dB PSNR, 1680 bits), **lower image:** MAC (37.34 dB PSNR, 1682 bits).

Again, both frames are transmitted at the same bit-rate. The model-aided coder yields a PSNR improvement of more than 2 dB.

Finally, let us take a closer look at the robustness of the proposed video codec. Figure 8 shows the model-based prediction of frame 18. As can be seen in the enlargement of the image region around the mouth, a model failure occurs that causes the black bar inside the mouth. Background and parts of the hair are also missing in the model frame. However, the rate-constrained coder control handles model failures robustly as illustrated by the selection mask in Fig. 8. In this figure all macroblocks are shown that are predicted from the model frame, while the macroblocks predicted from the reconstructed frame are grey.

6. 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. 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. In

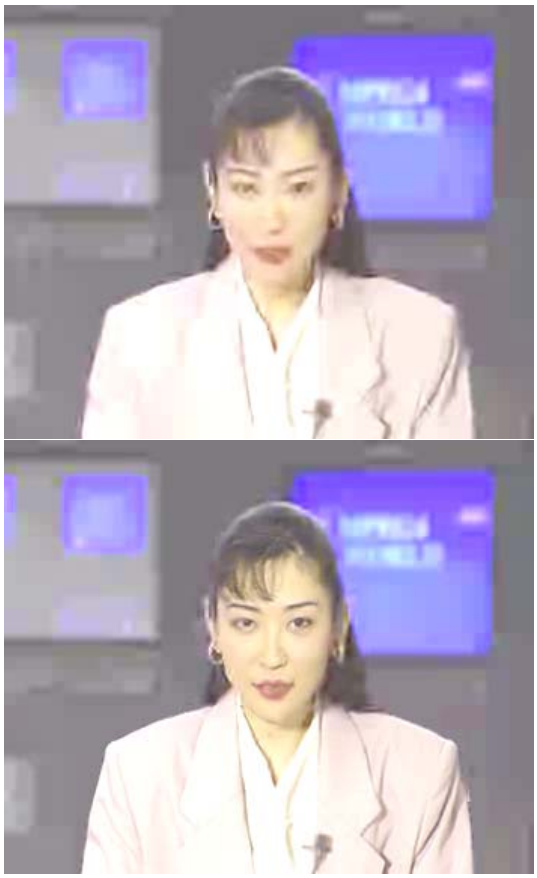

Figure 7: Frame 150 of the *Akiyo* sequence coded at the same bit-rate using the TMN-10 and the MAC, **upper image:** TMN-10 (31.08 dB PSNR, 720 bits), **lower image:** MAC (33.19 dB PSNR, 725 bits).

our scheme, two frames are utilized for prediction where one frame is the previous decoded one and the other frame is provided by a model-based coder. This yields increased robustness to estimation errors in the model-based coder, since rate-distortion optimization is employed.

7. REFERENCES

[1] W. J. Welsh, S. Searsby, and J. B. Waite, "Model-based image coding", *British Telecom Technology Journal*, vol. 8, no. 3, pp. 94–106, Jul. 1990.

[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.

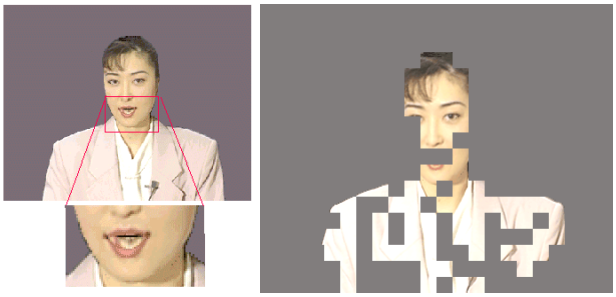

Figure 8: **left:** model frame 18 of the sequence *Akiyo* with enlargement of the image region around the mouth, **right:** corresponding selection mask showing those macroblocks that have been selected from the model frame.

[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] M. Kampmann and J. Ostermann, "Automatic adaptation of a face model in a layered coder with an object-based analysis-synthesis layer and a knowledge-based layer", *Signal Processing: Image Communication*, vol. 9, no. 3, pp. 201–220, Mar. 1997.

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

[9] M. Rydfalk, *CANDIDE: A Parameterized Face*, PhD thesis, Linköping University, 1978, LiTH-ISY-I-0866.

[10] G. Greiner and H. P. Seidel, "Splines in computer graphics: Polar forms and triangular B-spline surfaces", *Eurographics*, 1993, State-of-the-Art-Report.

[11] *ISO/IEC 14496-2, Coding of Audio-Visual Objects: Visual (MPEG-4 video), Committee Draft, Document N1902*, Oct. 1997.

[12] 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, June 1993.

[13] T. Wiegand, M. Lightstone, D. Mukherjee, T. G. Campbell, and S. K. Mitra, "Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 2, pp. 182–190, Apr. 1996.

[14] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression", *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, Nov. 1998.