LONG-TERM MEMORY PREDICTION USING AFFINE MOTION COMPENSATION

Thomas Wiegand, Eckehard Steinbach, and Bernd Girod

Telecommunications Laboratory
University of Erlangen-Nuremberg
Cauerstrasse 7/NT, D-91058 Erlangen, Germany
email: [wiegand|steinb|girod]@nt.e-technik.uni-erlangen.de

ABSTRACT

Long-term memory prediction extends motion compensation from the previous frame to several past frames with the result of increased coding efficiency. In this paper we demonstrate that combining long-term memory prediction with affine motion compensation leads to further coding gains. For that, various affine motion parameter sets are estimated between frames in the long-term memory buffer and the current frame. Motion compensation is conducted using standard block matching in the multiple reference frame buffer. The picture reference and the affine motion parameters are transmitted as side information. The technique is embedded into a hybrid video coder mainly following the H.263 standard. The coder control employs Lagrangian optimization for the motion estimation and macroblock mode decision. Significant bit-rate savings between 20 and 50 % are achieved for the sequences tested over TMN-10, the test model of H.263+. These bit-rate savings correspond to gains in PSNR between 0.8 and 3 dB.

1. INTRODUCTION

The most successful class of today's video compression designs are called hybrid codecs. The concept of block-based motion-compensated prediction (MCP) is prevalent in all these coding schemes [1]. The achievable MCP performance can be increased by reducing the size of the motion-compensated blocks [2]. This is taken into account, for example, by the INTER-4V mode of H.263 [3] where instead of a block of size 16×16 luminance pixels four blocks of size 8×8 luminance pixels are employed. However, the bit-rate must be assigned carefully to the motion vectors of these smaller blocks. Therefore, rate-constrained motion estimation is often employed yielding improved compression efficiency [2, 4, 1]. In rate-constrained motion estimation, a Lagrangian cost function $J = D + \lambda R$ is minimized, where distortion D is weighted against rate R using a Lagrange multiplier λ. Moreover, also the macroblock mode decision should be based on Lagrangian optimization techniques [5]. These facts have been recognized within the ITU-T/SG16/Q15 group when adopting TMN-10, the test model of the H.263 standard [6] which is based on Lagrangian optimization techniques [7].

Long-term MCP increases the efficiency of video compression schemes by utilizing several past frames that are assembled in a multi-frame buffer. This buffer is simultaneously maintained at encoder and decoder. Block-based MCP is performed using motion vectors that consist of a spatial displacement and a picture reference to address a block in the multi-frame buffer. Rate-constrained motion estimation is employed to control the bit-rate of the motion data. In [8], bit-rate savings around 15 % are reported when comparing a 10 frame long-term memory video coder to TMN-10, the test model of H.263. The ITU-T/SG16/Q15 group has decided to adopt this feature as an Annex to the H.263 standard [9].

Although, for long-term memory prediction, the motion model is extended to exploit long-term dependencies in the video sequence, the motion model remains translational. However, independently moving objects in combination with camera motion and focal length change lead to a complicated motion vector field which may not be efficiently approximated by variable block size MCP. With an increasing time interval between video frames, this effect is further enhanced since more complex motion is likely to occur. An example for a highly-optimized affine MCP scheme is presented by Karczewicz et al. in [10] that stimulated our work on that subject. In [10], an affine motion model is assigned to regions that consist of a set of connected blocks of size 8×8 pixels. Because of the inter-connectivity of these regions, a segmentation approach is proposed in [10] where the shape of the regions and the assignment of the motion models is optimized iteratively.

In this paper, we present a video codec that incorporates affine motion compensation without explicit segmentation into connected regions. For that, we extend the long-term memory coder in [8] by warping additional reference frames using affine parameters that correspond to motion clusters in the scene. The idea of warping reference frames has been published by the authors already in [11] where multiple warped versions of the previously decoded frame are employed for multi-frame block-based MCP. Note that in [12], a very similar approach has appeared that was developed independently to our work in [11]. Our scheme in [11] has been extended by a more efficient affine motion estimation algorithm that improves motion compensation performance and computational complexity. These extensions are mainly described in [13]. Additionally, the scheme in [11, 13] is extended by long-term memory prediction in that warped frames and block-based motion compensation can be conducted by referencing several past frames This is the subject of this paper.

0-7803-5467-2/99/ $ 10.00 © 1999 IEEE

2. AFFINE LONG-TERM MEMORY MOTION-COMPENSATED PREDICTION

The long-term memory coder is extended by warping additional reference frames using affine parameters. Given these reference frames and the long-term memory buffer, rate-constrained multi-frame block-based encoding algorithms are employed similar to [8]. The architecture of the multi-frame affine motion-compensated predictor is depicted in Fig. 1. It shows a video structure which uses $M = K + N$

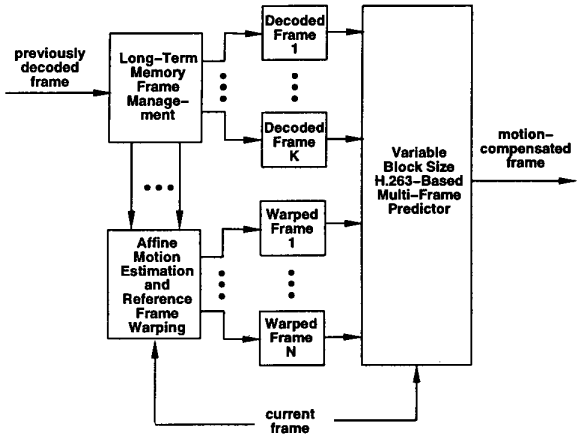

Figure 1: Architecture of the multi-frame affine motion-compensated predictor.

($M \geq 1$) picture memories for MCP. The M picture memories comprise two sets:

- K past decoded frames and
- N warped versions of past decoded frames.

The long-term memory frame management unit assembles K past decoded frames in a sliding window fashion, i.e., the oldest of the K frames is replaced by the most recent frame in the buffer [8].

The additional N reference frames are determined using the following two steps.

A Estimation of N affine motion parameter sets using the K previous frames and the current frame.

B Affine warpings of N additional reference frames.

Finally, the H.263-based multi-frame predictor conducts block-based MCP using all $M = K + N$ frames producing a motion-compensated frame. This motion-compensated frame is then used in a standard hybrid DCT video coder [1, 3]. The effective number of reference frames $M^* \leq M$ is determined by evaluating the rate-distortion efficiency in terms of Lagrangian costs for each reference frame. The chosen reference frames are indicated in the header of each picture. Hence, the picture reference parameters are encoded accordingly.

Relating our approach to region-based coding with affine motion models, e.g. see [10], we note that we also transmit various motion parameter sets and that the various "regions" associated with these motion parameter sets are indicated by the picture reference parameter. But, the "regions" in our scheme do not have to be connected. They are only restricted by the granularity of the variable block-size segmentation of the block-based video codec. Furthermore, each block belonging to a "region" may have an individual spatial displacement vector adapting the motion-compensated signal to the local statistics. This is beneficial if the motion exhibited in the scene cannot be compensated by a few affine motion models. In addition, if the video scene does not lend itself to a description by affine motion models, the coder drops into its *fall-back mode* which is block-based motion compensation using the previously decoded frame only.

Comparing the method of motion compensation using warped reference pictures to long-term memory MCP [8] we can draw the following conclusions. Both methods increase the probability of finding a good match for the block-based motion search at the cost of transmitting the reference frame index as side information. But, keeping the past decoded frames as reference pictures improves the rate-distortion performance only if there are repetitions in the scene. Hence if this assumption does not hold, for example when a scene cut occurs, long-term memory MCP is no longer efficient and the video coder drops into its *fall-back mode*.

Note that in a well-designed video codec, the most efficient concepts should be combined in such a way that their utility can be adapted to the source signal without significant overhead bit-rate. More precisely, if the overhead bit-rate needed to signal an encoding strategy is larger than the bit-rate savings achieved, the strategy should be removed from the video coding syntax. This idea is incorporated into the design strategy for many video codecs. Hence, the proposed video codec enables the utilization of variable block-size coding, long-term memory prediction and affine motion compensation. The use of the various motion compensation schemes can be signaled with very little overhead as described in the next section. The proposed motion compensation scheme is integrated into H.263+ demonstrating superior rate-distortion performance compared to TMN-10. As shown in the paper, both parts of the affine long-term memory prediction scheme contribute to the overall gain.

3. VIDEO SYNTAX

The affine long-term memory motion-compensated predictor is integrated into an H.263+-based video coding scheme. Our motivation for that is: (*i*) the algorithm is well defined and state-of-the-art [3], (*ii*) the highly optimized test model of the H.263+ standard, TMN-10, can be used as reference for comparison.

Motion estimation and compensation are performed using block-based multi-frame prediction mainly following the syntax and specifications of H.263+. The H.263+ syntax is modified in that changes are made to the inter-prediction modes INTER, INTER-4V, and UNCODED. 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 with one of the four 8×8 motion vectors. The syntax mainly follows the proposed Annex for long-term memory prediction in [9].

The parameters for the M^* chosen reference frames are

transmitted in the header of each picture. First, their actual number using a variable length code is signaled. Then, the temporal reference of each reference frame is encoded followed by a bit indicating whether the frame is warped or not. If that bit indicates a warped frame, a six parameter affine motion model is transmitted. The motion model employed in our codec is an orthonormalized version of the well known affine model. We adopt the orthonormalization of this model as derived in [10] to make the reconstructed motion model less sensitive to quantization of the model coefficients. A uniform scalar quantizer followed by entropy coding is used in our implementation. For more details on syntax please see [14]. Please note that the syntax specifying the actual number of reference frames allows the adaptation of the encoded bit-stream to the source signal on a frame basis without incurring much overhead.

4. RATE-CONSTRAINED CODER CONTROL

Rate-constrained reference frame generation and motion compensation proceed in three steps:

1 Estimate N affine motion parameter sets and warp the corresponding reference frames.

2 Conduct multi-frame block-based motion estimation on the reference frames.

3 Determine the number of affine motion parameter sets that are efficient in terms of rate-distortion performance.

The estimation of the N affine long-term memory motion parameter sets proceeds as follows. The current frame is partitioned into N blocks. In our implementation, N is typically chosen as 32, and the frame is partioned into 8×4 blocks of size 22×36 pixels. For each block, a half-pel accurate translational motion vector is estimated relative to each of the K frames in the long-term memory buffer via block matching in a search range of ± 16 pixels horizontally and vertically. Each of these K motion vectors is used as initialization of a gradient-based affine motion estimation procedure. Finally, that motion model of the K estimated motion models is chosen which minimizes the average distortion between the original pixels and the affine warped pixels in that block. Affine warping is performed using cubic spline interpolation between pixels [10]. Then, complete reference frames are warped each corresponding to an affine motion model. For more details on affine motion estimation and reference frame warping please refer to [13].

At this point it is important to note that the multi-frame buffer is filled with the K most recent frames and N warped frames yielding M reference frames. In order to produce the motion-compensated frame, we conduct multi-frame block-based motion compensation similar to H.263. That is, half-pel accurate motion vectors $\mathbf{v} = (v_x, v_y, \Delta)^T$ are applied to compensate blocks of size 16×16 pixels (UNCODED or INTER mode) or blocks of size 8×8 pixels (INTER-4V mode). More precisely, block-based motion estimation is conducted by minimizing

$$D_{DFD}(\mathcal{B}, \mathbf{v}) + \lambda_{MODE} \cdot R(\mathbf{v}) \qquad (1)$$

where the distortion $D_{DFD}(\mathcal{B}, \mathbf{v})$ for the block $\mathcal{B}$ between the original frame o and the reconstructed frame s_Δ that is indexed via the picture reference parameter Δ is computed as

$$D_{DFD}(\mathcal{B}, \mathbf{v}) = \sum_{x,y \in \mathcal{B}} (o[x, y] - s_\Delta[x + v_x, y + v_y])^2, \qquad (2)$$

$R(\mathbf{v})$ is the bit-rate associated with the motion vector including spatial displacement and picture reference parameter.

Given the motion vectors, the macroblock modes are chosen. Again, we employ a rate-constrained decision scheme where a Lagrangian cost function is minimized for each macroblock [5]

$$D_{REC}(\mathcal{B}, h, \mathbf{v}, c) + \lambda_{MODE} \cdot R_{INTER}(\mathcal{B}, h, \mathbf{v}, c), \qquad (3)$$

Here, the distortion after reconstruction D_{REC} measured as sum of squared differences between pixels of the block $\mathcal{B}$ in the original and reconstructed frame is weighted against bit-rate R_{REC} using the Lagrange multiplier λ_{MODE}. R_{REC} is the bit-rate that is needed to transmit and reconstruct a particular mode, including the macroblock header h, motion information $\mathbf{v}$ and DCT coefficients c. The mode decision determines whether to code each macroblock using the H.263 modes INTER, UNCODED, INTER-4V, and INTRA [3]. Following [1], the Lagrange multiplier for the mode decision is chosen as $\lambda_{MODE} = 0.85 \cdot Q^2$, with Q being the DCT quantizer value, i.e., half the quantizer step size [3]. The Lagrange multiplier used in the motion estimation is chosen as $\lambda_{MOTION} = \lambda_{MODE}$.

However, at this point, there is still an open problem about the efficient combination of motion vectors, macroblock modes and reference frames. Because of the dependency of the various parameters, we first pre-compute and store the Lagrangian costs for each combination of macroblock modes and reference frames in an array and then search a locally optimal solution given this array of data. The precomputation step proceeds as:

1 Determine the rate-distortion cost for the UNCODED mode for each of the M frames by computing its Lagrangian cost

$$D_{DFD}(\mathcal{B}, (0, 0, \Delta)^T) + \lambda_{MODE} \cdot R_{UNCODED}, \qquad (4)$$

with D_{DFD} being computed using (2) and $R_{UNCODED}$ being the bit-rate for the UNCODED mode.

2 Determine the rate-distortion cost for the INTER mode for each of the M frames by

- **A** Estimation of the best integer-pel accurate 16x16 block motion vector $\mathbf{v}$ in terms of its associated Lagrangian costs using (1) and (2).
- **B** Half-pel refinement of this integer-pel motion vector.
- **C** Computation of the macroblock Lagrangian cost for that half-pel refined motion vector including distortion and rate after DCT.

3 Determine Lagrangian cost for the INTRA mode.

The costs associated with the INTER-4V macroblock mode are computed later in order to reduce the computational burden.

Given the array of Lagrangian costs for the three modes UNCODED, INTER and INTRA, the number of affine motion parameter sets that are efficient in terms of rate-distortion performance are determined by the following algorithm.

1 Sort the reference frames according to the frequency of their selection.

2 Starting with the least popular reference frame, test the utility of each reference frame by

A Computing its best replacement block by block among the more popular frames in terms of rate-distortion costs.

B If the costs for transmitting the reference frame parameters exceed the cost of using the replacements for this frame, remove the frame, otherwise keep it.

After having determined the number of optimum frames M^* in the multiple reference frame buffer, the rate-distortion cost of the INTER-4V mode is considered and the selected parameters are encoded. Note that other optimization algorithms like a dynamic programming approach similar to [5] could be employed. However, due to the affine warping approach, the dependency between the various motion vectors is highly reduced. In fact, most of the vectors that point into warped reference frames are small. Therefore we can restrict the integer-pel search range when referencing a warped frame to ± 2 pixels horizontally and vertically without incurring significant degradation. In contrast to that, when searching decoded frames, the full range of ± 16 spatially displaced pixels is considered.

5. EXPERIMENTAL RESULTS

Experiments were conducted using the QCIF test sequences *Foreman*[1], *Mobile & Calendar, News*, and *Silent Voice*. A total of 100 frames of the sequences *Foreman, Mobile & Calendar*, and *News* are encoded at 10 Hz while 150 frames of the sequence *Silent Voice* are encoded at 15 Hz temporal sampling rate. Rate-distortion plots are generated by varying the DCT quantizer over values $4, 5, 7, 10, 15$, and 25. Bit-streams are generated that are decodable producing the same PSNR values as at the encoder. The data of the first intra-coded frame are excluded from the results.

Figures 2-5 show the average PSNR from reconstructed frames produced by the TMN-10 codec, the long-term memory prediction codec using 10 frames with and without affine warping vs. overall bit-rate. The size of the long-term memory is selected as $K = 10$ frames. The number of estimated affine motion models is $N = 32$. The three codecs differ only in the MCP part. The encoding strategy for all coders compared follows TMN-10, see [1]. Various observations can be made:

- Long-term memory MCP with 10 frames and without affine warping is always better than TMN-10. Bit-rate savings up to 15 % are achieved for the sequences tested.
- Long-term memory MCP with 10 frames and affine warping is always better than the case without affine warping. Bit-rate savings due to affine long-term memory MCP are up to 35 % (see Fig. 3) compared to long-term memory MCP without affine motion compensation.
- Long-term memory MCP with 10 frames and affine warping is superior compared to TMN-10 in terms of rate-distortion performance. Bit-rate savings up to 50 % can be achieved that correspond to PSNR gains of 3 dB (see Fig. 3).
- The gains for sequences with a large amount of motion (like *Mobile & Calendar*) are larger than those for low motion sequences (like *Silent Voice*).

[1] The short version of the *Foreman* sequence is used that was distributed within MPEG-4.

6. CONCLUDING REMARKS

The extension of long-term memory prediction by affine reference picture warping yields a superior video coding scheme in terms of rate-distortion performance. Significant coding gains between 20 and 50 % are obtained for the sequences tested when comparing to TMN-10, the test model of H.263. These bit-rate savings correspond to gains in PSNR between 0.8 and 3 dB. Bit-streams and a decoder can be down-loaded via anonymous ftp:

```
ftp ftp.nt.e-technik.uni-erlangen.de
cd pub/wiegand/MRPW
```

The affine long-term memory codec have been submitted as a proposal to ITU-T/SG16/Q15 for potential inclusion into H.263 Version 3 and the future video compression standard H.26L [14].

7. REFERENCES

[1] 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.

[2] G. J. Sullivan and R. L. Baker, "Rate-Distortion Optimized Motion Compensation for Video Compression Using Fixed or Variable Size Blocks", in *Proc. GLOBECOM'91*, 1991, pp. 85–90.

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

[4] B. Girod, "Rate-Constrained Motion Estimation", in *Proceedings of the SPIE Conference on Visual Communications and Image Processing*, Chicago, USA, Sept. 1994, pp. 1026–1034.

[5] 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.

[6] ITU-T/SG16/Q15-D-65, "Video Codec Test Model, Near Term, Version 10 (TMN-10), Draft 1", Download via anonymous ftp to: standard.pictel.com/videosite/9804_Tam/q15d65d1.doc, Apr. 1998.

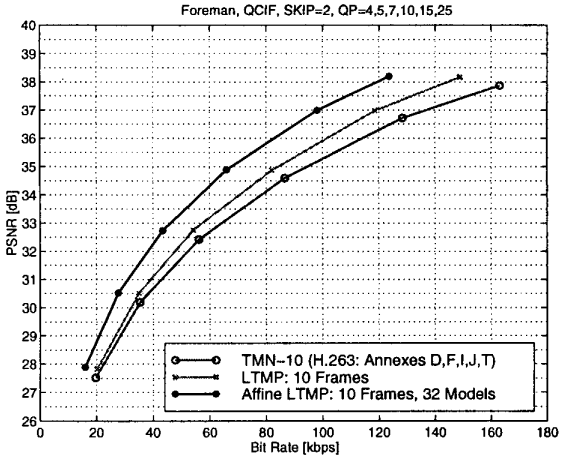

Figure 2: PSNR vs. overall bit-rate for the QCIF sequence *Foreman*.

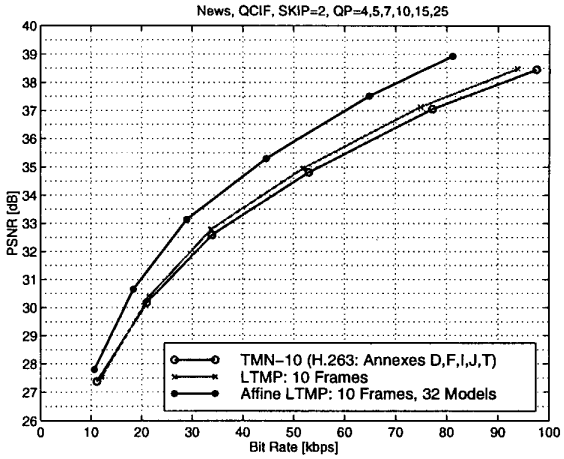

Figure 4: PSNR vs. overall bit-rate for the QCIF sequence *News*.

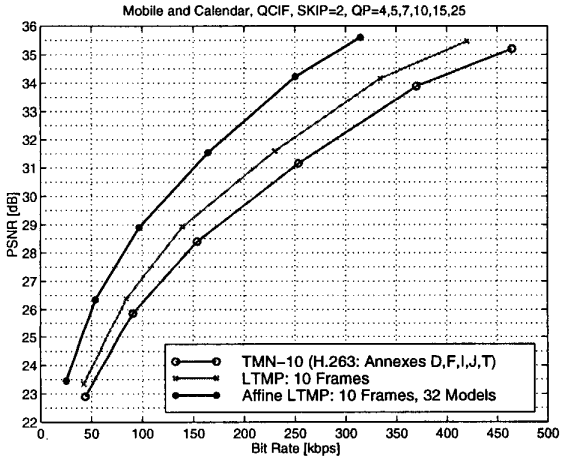

Figure 3: PSNR vs. overall bit-rate for the QCIF sequence *Mobile & Calendar*.

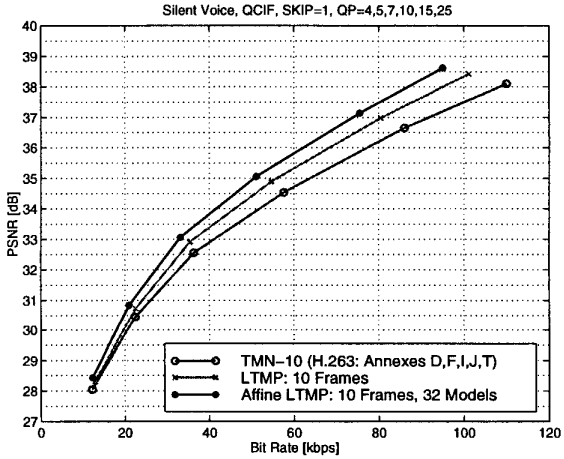

Figure 5: PSNR vs. overall bit-rate for the QCIF sequence *Silent Voice*.

[7] ITU-T/SG16/Q15-D-13, T. Wiegand and B. Andrews, "An Improved H.263-Codec Using Rate-Distortion Optimization", Download via anonymous ftp to: standard.pictel.com/video-site/9804_Tam/q15d13.doc, Apr. 1998.

[8] T. Wiegand, X. Zhang, and B. Girod, "Long-Term Memory Motion-Compensated Prediction", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 70–84, Feb. 1999.

[9] ITU-T/SG16/Q15-G-18, T. Wiegand, N. Färber, B. Girod, and B. Andrews, "Proposed Draft for Annex U on Enhanced Reference Picture Selection", Download via anonymous ftp to: standard.pictel.com/video-site/9902_Mon/q15g18.doc, Feb. 1999.

[10] M. Karczewicz, J. Niewęgłowski, and P. Haavisto, "Video Coding Using Motion Compensation with Polynomial Motion Vector Fields", *Signal Processing: Image Communication*, vol. 10, pp. 63–91, 1997.

[11] T. Wiegand, E. Steinbach, A. Stensrud, and B. Girod, "Multiple Reference Picture Coding using Polynomial Motion Models", in *Proceedings of the SPIE Conference on Visual Communications and Image Processing*, San Jose, USA, Feb. 1998, pp. 134–145.

[12] D. Lauzon and E. Dubois, "Representation and Estimation of Motion Using a Dictionary of Models", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 2585–2588.

[13] E. Steinbach, T. Wiegand, and B. Girod, "Using Multiple Global Motion Models for Improved Block-Based Video Coding", in *Proceedings of the IEEE International Conference on Image Processing*, Kobe, Japan, Oct. 1999.

[14] ITU-T/SG16/Q15-G-21, T. Wiegand, E. Steinbach, B. Girod, and B. Andrews, "Video Coding Using Long-Term Memory and Affine Motion Compensation", Download via anonymous ftp to: standard.pictel.com/video-site/9902_Mon/q15g21.doc, Feb. 1999.