THE MOTION TRANSFORM: A NEW MOTION COMPENSATION TECHNIQUE

Robert M. Armitano and Dinei A. F. Florêncio and Ronald W. Schafer

Digital Signal Processing Laboratory
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332

armi@eedsp.gatech.edu floren@eedsp.gatech.edu rws@eedsp.gatech.edu

ABSTRACT

Motion estimation plays an important role in video coding schemes by removing temporal redundancies that exist in image sequences. Motion information that is required in motion compensated prediction is transmitted as side information, as in forward motion estimation, or can be computed at the receiver in backward motion estimation. The former method has the advantage of more accurate prediction but, suffers in that the motion information is transmitted as side information. We introduce the Motion Transform (MT), a new motion estimation and compensation technique that does not require the transmission of motion vectors yet, performs well under a variety of conditions. In the proposed method, motion information is obtained in a bottom-up hierarchical fashion using a two-dimensional wavelet-like decomposition.

1. INTRODUCTION

Motion estimation plays an important role in video coding schemes by reducing temporal redundancy in an image sequence. In video coding standards (e.g. MPEGI, MPEGII and H.263 [1, 2, 3].), a portion of the available channel bandwidth is occupied by this motion side information. To circumvent the overhead requirements of motion compensated predictive coding, the Motion Transform (MT) is proposed. The MT is a motion estimation and compensation scheme that does not require the transmission of motion vectors. Motion information is obtained in a coarse-to-fine hierarchical decomposition using non-linear non-expansive filterbanks [4].

In the proposed method, motion estimation is performed between frames at different resolution. Initially the lowest resolution representation of the frame difference is transmitted. Using the reconstructed lowest resolution frame, motion vectors are computed for the next level in the image hierarchy. Any previously transmitted frame can be used as a reference. This motion information is used to predict the frame at the current level of resolution. By adding the predicted frame to the motion compensated residual the frame is reconstructed. Refinement of increasingly higher resolution motion vectors is possible as the decoder works up through the image hierarchy. Motion estimation is performed at each resolution level to provide the most accurate predicted frame, until the full resolution image is recovered.

Although technically speaking the proposed method can be classified as backward motion estimation, it performs as well as the block matching algorithms (BMA) without motion vector transmission. In addition, by reconstructing the image in an hierarchical fashion, samples on both sides of the pixel being evaluated are provided unlike current backward motion estimation algorithms. Depending on design constraints, block based or pixel based motion estimation routines can be employed during motion estimation.

In Section 2 we review motion compensated prediction as well as forward and backward motion estimation methods. The motion transform is a backward motion estimation technique and is presented in Section 3. Even though the MT is classified as a backward motion estimation method its performance is as good as forward motion estimation algorithms as shown in Section 4. In Section 5 the conclusions are presented.

2. MOTION COMPENSATED PREDICTION

In video coding, motion information is used to displace pixels in a reference frame to form a predicted frame. The predicted frame is subtracted from the original frame reducing temporal redundancies in the signal and lowering the signal's bandwidth requirements. This technique is known as motion compensated prediction and is used in the majority of commercial video coders today. The accuracy of the displacement vector dictates the quality of the predicted frame and the energy in the motion compensated residual. Currently there are two main methods used for motion estimation: (1) forward motion estimation and (2) backward motion estimation.

Forward motion estimation, bases the motion estimation on the current frame plus a previously encoded (and transmitted) frame. The immediate drawback is the need to transmit the motion vectors, which consumes a portion of the available bitrate. The most common technique in this class is the Block Matching Algorithm (BMA) that is used

This work is supported in part under a grant from the Hewlett-Packard Company, in part under the Joint Services Electronics Program, Contract DAAH-04-93-G-0027, and by CNPq (Brazil) under contract 200.246-90/9.

0-7803-3192-3/96 $5.00©1996 IEEE

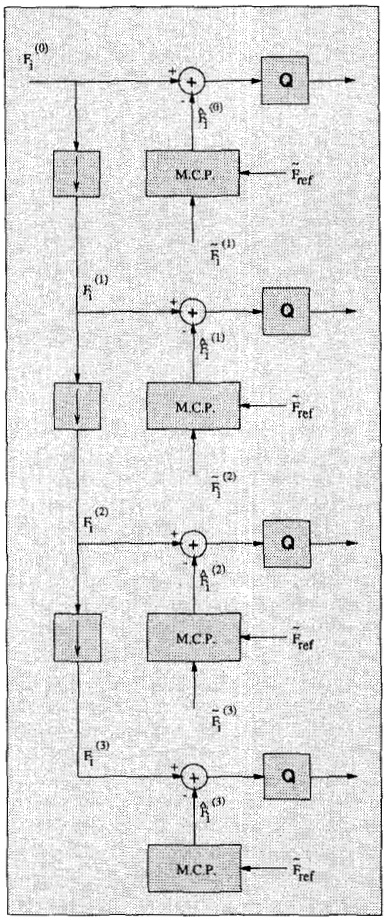

Figure 1: **Motion Transform Encoder.**

in the MPEGI, MPEGII and H.263 video coding standards.

When using FME, motion information is transmitted as side information. There is a tradeoff between the motion vector overhead and the bandwidth required to transmit the residual signal. Motion vectors that faithfully represent motion in a scene provide a more accurate prediction, thereby reducing the entropy of the residual. However, more bits are required to transmit this accurate motion field.

Backward Motion Estimation, bases motion estimation solely on information available at both encoder and decoder, and does not require transmission of motion vectors. In a BME-type implementation, two previous frames are used to estimate the motion for the current frame. The motion estimate is not as good as the FME BMA but is superior to the frame difference.

Reliability of the motion vectors is the main issue when using BME. The fact that the estimate is based only on past frames makes the method very sensitive to sudden changes

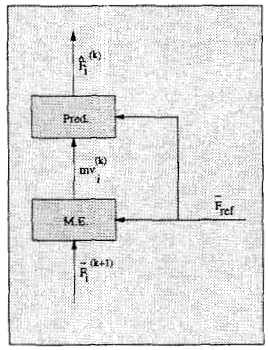

Figure 2: **Motion Compensation Block.**

in movements and to complex movements such as those generally taken by human body parts (mouth, face, arms, etc). In pel-recursive motion estimators this disadvantage is off-set by providing one motion vector for each pixel in the image. The pel-based algorithm however, can only make use of pixels that have already been transmitted for the current frame. Due to the progressive nature of transmission, the pixels from the current frame are only available at one side and above of the pixel whose motion is being estimated [5].

3. THE MOTION TRANSFORM

The MT uses a wavelet-like hierarchical decomposition using non-linear filterbanks to represent the input signal, as shown in Fig. 1. At time i, F_i^0 is the full resolution image. At each stage of the decomposition a quincunx decimator is used to sub-sampled the input signal by a factor of two,

$$F^{*(k+1)}[n,m] = F^{(k)}[2n + (n \bmod 2), m] \qquad (1)$$
$$F^{(k+1)}[n,m] = F^{*(k+1)}[n, 2m]$$

where $F^{(k+1)}$ is the lower resolution image and $F^{*(k+1)}$ is an intermediate signal, with a quincunx sampling grid. The spatial decomposition is performed repeatedly until the lowest resolution image is obtained.

To code the lowest resolution image F_i^K an estimate, $\hat{F}_i^K$, is obtained using motion compensated prediction. Motion information can be obtained using linear predictive coding or in the example in this paper, the lowest resolution motion vector is set to zero, and the predicted frame $\hat{F}_i^K = \tilde{F}_{ref}^K$. In other words, the frame difference is coded and sent to the receiver.

For subsequent levels in the hierarchy a lower resolution image is always available for motion estimation. At level k, decoded image $\tilde{F}_i^{k+1}$, is used as the current frame for motion estimation, as shown in Fig. 2. Any frame that was previously transmitted can be used as the reference frame, $\hat{F}_{ref}^0$. Displacement vectors are found for blocks of pixels, $\mathbf{b}^k$, at resolution k. Any motion estimation technique can be used

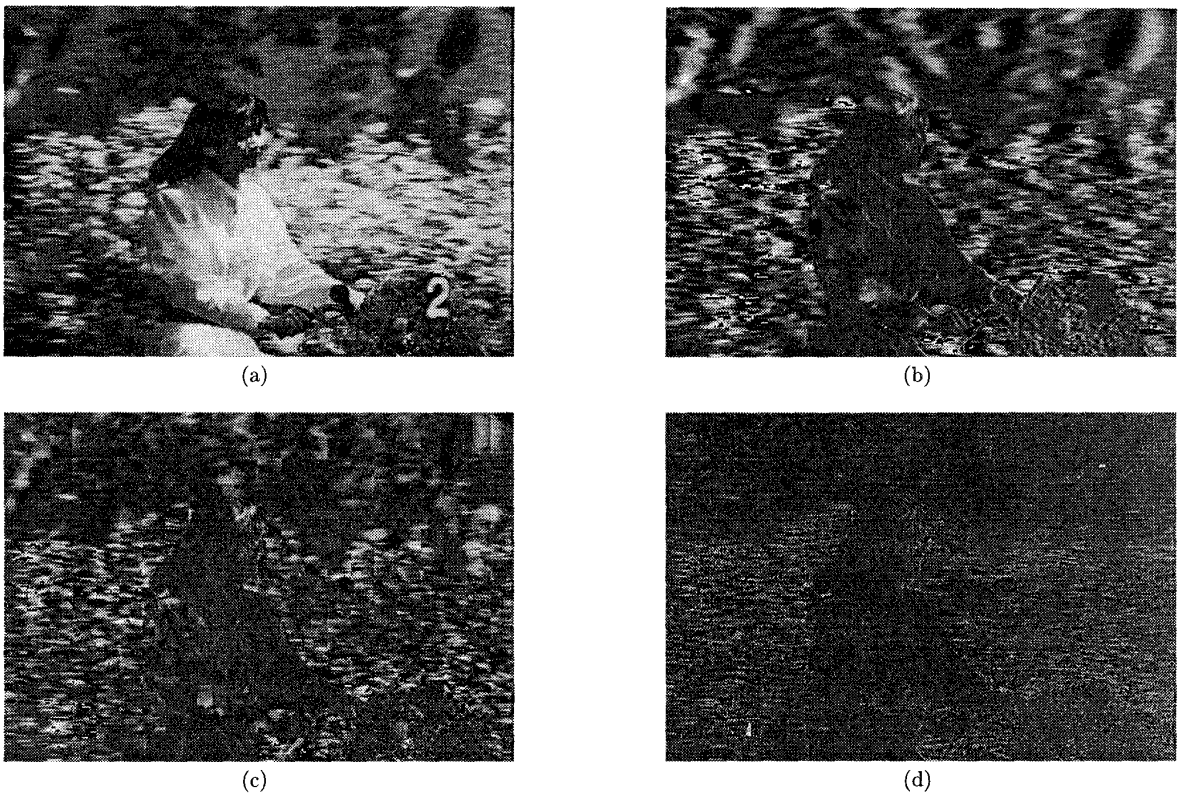

Figure 4: **Motion Transform Example**

Frame 52 of the CycleGirl *sequence is used to illustrate the MT. The original frame is shown in (a). The frame difference is shown in (b). The motion compensated residual for the block matching algorithm using an* (8×8) *search region is shown in (c). The Motion Transform is shown in (d) with a* (8×8) *search region.*

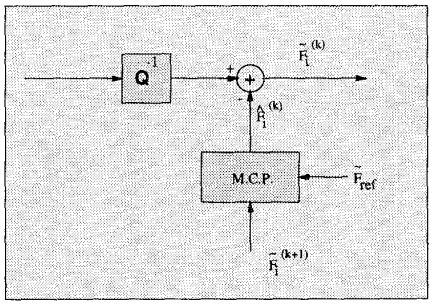

Figure 3: **Motion Transform Decoder.**

to find the best match and any size block can be used at each stage in the hierarchy. Once an appropriate match is found, b^k_{ref} is composited into the predicted frame. The predicted frame $\hat{F}^k_i$ is constructed on a block-by-block basis using displaced subsampled blocks from $\tilde{F}_{ref}$. The predicted signal, $\hat{F}^k_i$, is then used to form the motion compensated residual signal which is transmitted to the receiver.

At the receiver the process is reversed as shown in Fig. 3. Technically the MT is a BME technique since it does not rely on motion vector side information but, since motion estimation can be performed using a lower resolution representation of the image, the MT performs as well as FME methods.

Some of the advantages of the proposed method:

- It does not require transmission of motion vector side information, as required with forward prediction techniques.
- It bases the estimation on the current and refined frame, as in the more efficient forward prediction BMA, instead of two past frames, as in backward prediction techniques.
- It uses pixels from every side, and not only from the upper (and left) pixels, as in pel-recursive methods.

- It does actual matching, in contrast to Wang and Clark [5], which uses only 1-pixel grey level difference as the motion matching criteria.
- Since it does not transmit side-information, all available information can be used, including small block sizes, pixel-by-pixel estimates or sub-pixel motion vector resolutions.

The disadvantages are:

- Motion information must be computed at both the encoder and decoder. This doubles the computational complexity of the video codec.
- At very-low-bit rates, the savings in not transmitting the motion vectors becomes more significant. Nonetheless, as quantization error increases, the reliability of the motion estimation based on these quantized samples also decreases.

4. RESULTS

Results for the MT are given for a three stage implementation, using block matching for motion estimation with (4×4) blocks at the lowest level of the hierarchy. Subsampling by two at each stage of the hierarchy was used for spatial decomposition. A DC-error compensation was performed on a pixel-wise basis with non-overlapping blocks for motion compensation.

The MT is shown for frame 52 in the CYCLEGIRL image sequence in Fig. 4.d. The residual energy is displayed for every component in the decomposition (for comparison purposes the individual residuals were composited up to the full resolution frame). The main objective of motion compensated prediction is to reduce the energy of the signal to be transmitted. Using the MT, we have obtained residual energies of 22.38 dB in this example, with a (8×8) search region. This is better than simple frame differencing, Fig. 4.b, with energy of 16.32 dB. The MT even compares favorably with the block matching algorithm, Fig. 4.c, with residual energy of 18.98 dB, (using a (16×16) block size, a (8×8) search region and full search matching). The full search BMA still requires motion vector transmission, while the MT does not.

In Fig. 5 the MT's performance is compared to the full search BMA for 95 frames in the CYCLEGIRL image sequence. A (8×8) search region was used in both cases. The MT outperforms the full search by a large margin.

5. CONCLUSIONS

The Motion Transform is a motion compensation technique that does not rely on the transmission of motion vector as side information. Using a non-linear non-expansive filter bank a hierarchical decomposition of the image allows coarse-to-fine motion compensation at the decoder without motion vector transmission. Displacement information at each stage of the hierarchy is obtained through motion estimation using the reconstructed lower resolution image as the current frame, while the reference frame is taken from the set of previously transmitted frames. The MT performs

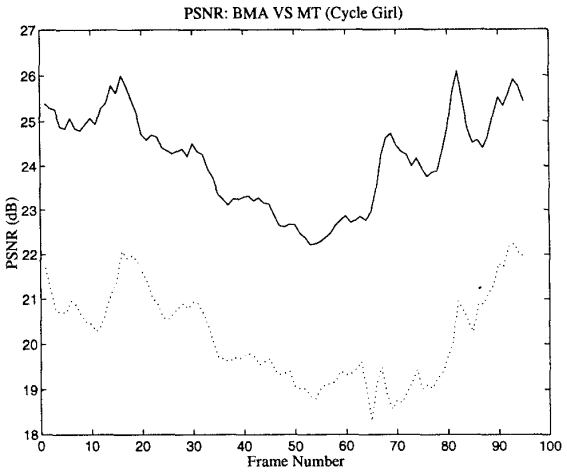

Figure 5: **PSNR Comparison.**
The full search block matching algorithm (dotted line) is compared to the MT (solid line) for the CYCLEGIRL *sequence.*

favorably in typical image sequences but, the motion information must be computed at the decoder as well as the encoder. In applications where asymmetrical coders are needed (e.g. broadcast environment) the MT use may be undesirable. The MT shows promise in applications able to tolerate symmetric computational loads, (e.g. teleconferencing) or in applications with strict channel bandwidth limitations, (e.g. the transmission of remote sensor imagery or video archiving).

6. REFERENCES

[1] V. Bhaskaran and K. Konstantinides, eds., *Image and Video Compression Standards.* Kluwer Academic Publishers, 1995.

[2] M. I. Sezan and R. L. Lagendijk, eds., *Motion Analysis and Image Sequence Processing.* Kluwer Academic Publishers, 1993.

[3] H. G. Musmann, P. Pirsch, and H.-J. Grallert, "Advances in picture coding," *Proceedings of the IEEE*, vol. 73, pp. 523–548, April 1985.

[4] D. A. F. Florêncio and R. W. Schafer, "Perfect reconstructing non-linear filter banks," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1996.

[5] Q. Wang and R. J. Clarke, "Motion estimation and compensation for image sequence coding," *Signal Processing*, vol. 4, pp. 161–174, 1992.