

MESH-BASED MOTION ESTIMATION AND COMPENSATION IN THE WAVELET DOMAIN USING A REDUNDANT TRANSFORM

Suxia Cui, Yonghui Wang, and James E. Fowler
*Department of Electrical and Computer Engineering
Engineering Research Center
Mississippi State University, Mississippi State, MS*

ABSTRACT

In this paper, a technique is presented that incorporates an irregular triangle mesh into wavelet-domain motion-estimation and motion-compensation using a shift-invariant redundant wavelet transform. Triangle vertices are identified by a simple correlation operator locating image edges in the wavelet subbands, while motion compensation takes place through an affine transformation mapping triangles from one frame to the next. The motion-compensated residual is downsampled to a non-redundant form which is then coded using any wavelet-based still-image coder. Experimental results indicate that the combined approach outperforms either technique applied separately; in addition, the proposed method outperforms a variety of motion-estimation and motion-compensation approaches operating in both the spatial and wavelet domains.

1. INTRODUCTION

Block-based motion estimation (ME) and motion compensation (MC) followed by a discrete cosine transform (DCT) is widely employed in modern video-compression systems and an integral part of standards such as H.263, MPEG-2, and MPEG-4. However, given the promising performance of wavelet-based still-image compression algorithms such as [1], there has recently been interest in deploying ME/MC within such algorithms to produce wavelet-based video coders. On another front, there have been a variety of proposals for employing geometries more general than square blocks to drive ME/MC. In this paper, we combine both of these recent developments. Specifically, we investigate the use of triangle-mesh based ME/MC in the wavelet domain.

The most straightforward way to replace the DCT with a discrete wavelet transform (DWT) in a typical video coder is to perform ME/MC in the spatial domain and to calculate a DWT on the resulting residual image. This simple approach suffers from blocking artifacts [2], which are exacerbated if the DWT is not block-based but rather the usual whole-image transform. The alternative paradigm would be to have ME/MC take place in the wavelet domain. However, the fact that the usual, critically sampled DWT used ubiquitously in image-compression efforts is shift variant greatly hinders the ME/MC process [3].

In this paper, we adopt the latter approach—wavelet-domain ME/MC. However, to overcome difficulties associated with the shift variance of traditional DWTs, we choose instead to perform ME/MC in the domain of a redundant transform. In essence, the redundant DWT (RDWT) [4] removes the downsampling operation from the traditional DWT to ensure shift invariance at the cost of a redundant, or overcomplete, representation.

The second key aspect of our approach is that we drive ME/MC with an irregular triangle mesh rather than the traditional block-

based structure. The motivation for mesh-based ME/MC is that a mesh structure can oftentimes better match the motion of objects in video. For example, highly detailed areas should be divided into many small irregularly shaped regions to be individually compensated, whereas larger ME/MC regions can suffice for areas with little detail. This fine-tuning of ME/MC is impossible in traditional block-based approaches since the size of the block is fixed. But in mesh-based approaches, such as triangle-mesh ME/MC [5], the regions are sized and shaped accordingly to the local level of detail in the image. Specifically, in triangle-mesh ME/MC, triangle vertices, or “control points,” are selected to track edges of objects in the image.

In this paper, we describe in detail our approach to triangle-based ME/MC in the RDWT wavelet domain and compare it empirically to several other ME/MC methods deployed in both the spatial and wavelet domains. Results indicate that our redundant-wavelet triangle-mesh (RWTM) method outperforms the other methods on both a fast-moving and a slow-moving video segment.

2. REDUNDANT-WAVELET TRIANGLE-MESH MOTION ESTIMATION AND MOTION COMPENSATION

The encoder of our RWTM video-coding system is depicted in Fig. 1. The input image is first transformed using a RDWT, and control points are identified in the previous frame by locating the most salient image edges. The motion of these control points from the previous frame to the current frame is estimated in the RDWT domain, and motion vectors are transmitted to the decoder to allow it to track control-point motion. MC is accomplished by first using a triangulation algorithm to generate triangle meshes from the control points in both the current and previous frames and then using affine transformations to predict, subband by subband, triangles in the current frame from triangles in the previous frame. Residing in the RDWT domain, the motion-compensated residual is itself redundant; consequently, it is downsampled before coding. The final encoding step consists of a wavelet-domain still-image coder; for the experiments below, we use SPIHT [1], but any wavelet-domain still-image coder would suffice.

At the decoder side, motion of the control points is tracked, and triangulations identical to those used in the encoder are produced. A reconstructed spatial-domain image is produced by inverting the still-image coding, adding on a subsampled RDWT-domain prediction, and inverting the DWT. Finally, a RDWT operation produces the reference-frame subbands for generating the prediction of the next-frame subbands in the RDWT domain. Below, we explore the various components of our proposed system in greater detail.

2.1. Redundant Wavelet Transform

The RDWT is an approximation to the continuous wavelet transform that removes the downsampling operation from the traditional critically sampled DWT to produce an overcomplete representation. The shift-variance characteristic of the DWT arises from its use of downsampling, while the RDWT is shift invariant since the spatial sampling rate is fixed across scale. As a result, the size of each subband in an RDWT is the exactly the same as that of the input signal. It turns out that, by appropriately subsampling each subband of an RDWT, one can produce exactly the same coefficients as does a critically sampled DWT applied to the same input signal. We note that the RDWT is also sometimes called the “algorithme à trous” or the “undecimated wavelet transform.” The reader is referred to [4] for greater detail on the RDWT, its implementations, and its relation to the critically sampled DWT.

2.2. Selection of Control Points

The choosing of proper control points is crucial to the accuracy of triangle-mesh ME. Typically, one wants control points to track salient image features (e.g., edges). The redundancy of the RDWT facilitates the identification of salient features in an image, especially image edges, since a simple correlation operation can easily accomplish edge identification [6]. Specifically, the direct multiplication of the RDWT coefficients at adjacent scales distinguishes important features from the background due to the fact that wavelet-coefficient magnitudes are correlated across scales. Coefficient-magnitude correlation is well known to exist in the usual critically sampled DWT also; however, the changing temporal sampling rate makes the calculation of an explicit correlation mask across scales much more difficult for the critically sampled DWT [6].

The correlation mask we propose consists of multiplying the high-low (HL) bands, the low-high (LH) bands, and the high-high (HH) bands together and combining the products,

$$\text{mask}(x, y) = \left| \prod_{j=J_0}^{J_1} HL^{(j)}(x, y) \right| + \left| \prod_{j=J_0}^{J_1} LH^{(j)}(x, y) \right| + \left| \prod_{j=J_0}^{J_1} HH^{(j)}(x, y) \right|,$$

where J_0 and J_1 are the starting and ending scales, respectively, of the correlation operation. We note that calculation of the correlation mask in this manner is possible due to the fact that each RDWT subband is the same size as the original image. Fig. 2 shows the correlation mask for the first frame of the sequence “Susie,” where we use the subbands from the two highest-frequency scales in the products above.

To identify controls point within the correlation mask, we divide the mask into equally sized blocks and select the N points with the largest mask value as candidate control points. Certain candidate points are then eliminated from consideration to reduce the number of controls points, and thus motion-vector information, needed for the frame. Specifically, a minimum distance between neighboring control points is imposed to avoid very small triangles. Additionally, candidate points are subjected to thresholding so that points corresponding to local maxima in a block that are not close to the subband’s global maximum are discarded. The threshold must be tailored to specific sequences for best performance; sequences with faster motion or smaller objects need more control points. Finally we note that control points that are equally spaced along the image border are added to the points chosen via the correlation mask in the image interior so that the meshed area covers the entire image.

2.3. Motion Estimation

Each control point identified via the correlation mask has an associated motion vector describing the movement of that control point from the previous frame to the current frame. These motion vectors are obtained by finding the best matching point in the current frame for each control point in the previous frame. This match is accomplished by calculating the absolute difference of a small block centered at the control point in the previous frame and blocks in a search window about the control-point location in the current frame, similar to the usual block-based ME process and identical to the technique used in [5] for spatial-domain triangle-mesh ME. However, because our ME takes place in the RDWT domain, for a given vector in the search window, we calculate absolute differences for all the subbands and sum them together to produce a cross-subband distortion. We choose the vector that minimizes this cross-subband distortion as the motion vector for the current control point. In order to maximize distortion performance, only the coefficients that will survive the subsequent RDWT-to-DWT domain downsampling operation are counted towards the cross-subband distortion calculation.

2.4. Triangulation and Affine Transformation

As in the spatial-domain triangle-mesh ME/MC of [5], after the control points are selected in the reference frame and their motion is tracked to the current frame, triangle meshes are computed using Delaunay triangulation. A single triangle mesh is used for all subbands in the RDWT of a frame, as depicted in Fig. 3 (only the HL subbands are shown); this is possible since each RDWT subband has the same size. MC proceeds by mapping each triangle in the current frame backwards to the previous frame using an affine six-parameter model as described in [7]; this affine mapping is performed for the triangles in each subband separately.

3. EXPERIMENTAL RESULTS

Experimental results use the 100-frame “Football” sequence and the 70-frame “Susie” sequence, both grayscale sequences with a spatial resolution of 352×240 pixels and a temporal sampling of 30 frames/sec. (noninterlaced). The first frame is intra-encoded (I-frame) while all subsequent frames use ME/MC (P-frames). All wavelet transforms (DWT and RDWT) use the Cohen-Daubechies-Feauveau 9-7 filter [8] with symmetric extension, and all ME/MC methods use integer-pixel accuracy. Since SPIHT, used as the core compression engine in all experiments, produces an embedded coding, each frame of the sequence is coded at exactly the specified target rate.

We compare our proposed RWTM technique to various prominent spatial-domain and wavelet-domain ME/MC algorithms. Average PSNR figures for fixed bit rate are tabulated in Table 1, and frame-by-frame PSNR profiles are shown in Figs. 4 and 5. In these results, “Spatial Block” refers to block-based ME/MC in the spatial domain, the traditional method employed in video-coding standards. “Spatial OBMC” is overlapped block ME/MC in the spatial domain [9]. “Spatial Mesh” is the irregular triangle-mesh ME/MC in the spatial domain [5]. All the preceding spatial-domain ME/MC approaches are followed by an entire-image DWT applied to the residual image and then SPIHT coding of the DWT coefficients. The “DWT Block” approach transforms the input image into the DWT domain and then applies the usual block-based ME/MC in DWT domain. “RDWT Block” transforms the input image into RDWT domain and then employs block-based ME/MC in RDWT domain [3, 10]. In these wavelet-domain techniques, SPIHT coding is applied to the MC residual.

4. CONCLUSIONS

The experimental results shown in Table 1 and Figs. 4 and 5 indicate that our proposed RWTM method, which combines the advantages of the wavelet-domain with irregular-mesh ME/MC, outperforms other ME/MC techniques operating in both the spatial and wavelet domains. In terms of average PSNR performance (Table 1), RWTM outperforms its nearest competitor (block-based ME/MC in the RDWT domain [3, 10]) by 0.4 dB for both the fast-motion “Football” and the slow-moving “Susie” sequences. It is interesting to note that our combination of triangle-mesh ME/MC and RDWT-based ME/MC outperforms either technique applied alone.

The success of our approach lies in that the shift invariance of the RDWT makes it an ideal candidate for the implementation of ME/MC in the wavelet domain. In fact, the RDWT has indeed been used previously for ME/MC in [3, 10], although it was not recognized by the authors as such. Instead, these previous approaches partitioned the RDWT coefficients into multiple, critically sampled subband pyramids according to all possible phase shifts. The ME procedure would “switch” between these subband pyramids, each similar to a critically sampled DWT, as the phase of the motion under consideration changed. The RDWT, on the otherhand, preserves the spatial coherence of the coefficients, thereby facilitating implementation of the affine transformation needed in mesh-based ME/MC. Complex indexing and interpolation between the multiple subband pyramids would be needed to do the same in the prior, partitioned implementations as each vertex of a given triangle may have motion of a different phase. Additionally, the RDWT permits easy identification of control points through a simple correlation operation whereas spatial-domain mesh-based techniques typically employ a more costly convolution operator to identify edge locations. We anticipate, thus, that the RDWT will play a vital role to the development of next-generation video-coding standards if these are to exploit the recent advances in wavelet-based still-image coding.

5. REFERENCES

- [1] A. Said and W. A. Pearlman, “A New, Fast, and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 3, pp. 243–250, June 1996.
- [2] G. V. der Auwera, A. Munteanu, G. Lafruit, and J. Cornelis, “Video Coding Based on Motion Estimation in the Wavelet Detail Images,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, May 1998, pp. 2801–2804.
- [3] H.-W. Park and H.-S. Kim, “Motion Estimation Using Low-Band-Shift Method for Wavelet-Based Moving-Picture Coding,” *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 577–587, April 2000.
- [4] M. J. Shensa, “The Discrete Wavelet Transform: Wedding the À Trous and Mallat Algorithms,” *IEEE Transactions on Signal Processing*, vol. 40, no. 10, pp. 2464–2482, October 1992.
- [5] M. Eckert, D. Ruiz, J. I. Ronda, and N. Garcia, “Evaluation of DWT and DCT for Irregular Mesh-based Motion Compensation in Predictive Video Coding,” in *Visual Communications and Image Processing*, K. N. Ngan, T. Sikora, and M.-T. Sun, Eds. Proc. SPIE 4067, June 2000, pp. 447–456.

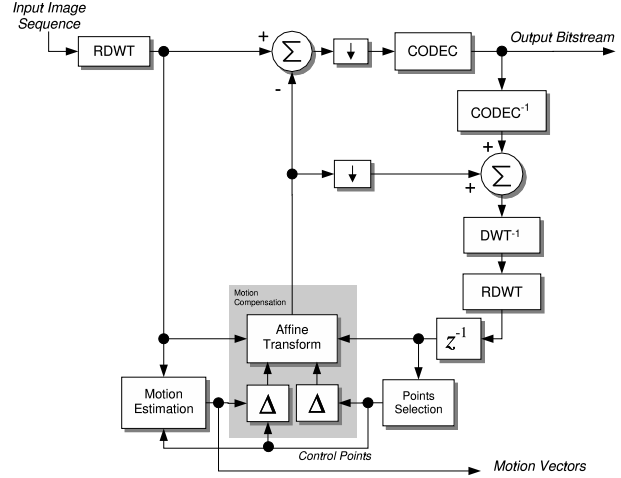


Figure 1: Block diagram of the RWTM video-coding system. z^{-1} = frame delay, \downarrow = subsampling from RDWT to DWT domains, Δ = triangulation. CODEC is any wavelet-based still-image coder.

Method	Football (0.5 bpp)	Susie (0.25 bpp)
Spatial Block	26.3	35.9
Spatial OBMC	27.4	34.6
Spatial Mesh	27.4	37.4
DWT Block	24.4	33.5
RDWT Block	27.7	37.4
RWTM	28.1	37.8

Table 1: Comparison of Average PSNR (dB)

- [6] Y. Xu, J. B. Weaver, D. Healy, Jr., and J. Lu, “Wavelet Transform Domain Filters: A Spatially Selective Noise Filtration Technique,” *IEEE Transactions on Image Processing*, vol. 3, no. 6, pp. 747–758, November 1994.
- [7] Y. Altunbasak, A. M. Tekalp, and G. Bozdagi, “Two-Dimensional Object-based Coding Using a Content-based Mesh and Affine Motion Parameterization,” in *Proceedings of the International Conference on Image Processing*, Washington, DC, October 1995, pp. 394–397.
- [8] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, “Image Coding Using Wavelet Transform,” *IEEE Transactions on Image Processing*, vol. 1, no. 2, pp. 205–220, April 1992.
- [9] S. A. Martucci, I. Sodagar, T. Chiang, and Y.-Q. Zhang, “A Zerotree Wavelet Video Coder,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 1, pp. 109–118, February 1997.
- [10] X. Li, L. Kerofsky, and S. Lei, “All-Phase Motion Compensated Prediction in the Wavelet Domain for High Performance Video Coding,” in *Proceedings of the International Conference on Image Processing*, Thessaloniki, Greece, October 2001, pp. 538–541.



Figure 2: Correlation mask for the first frame of "Susie".

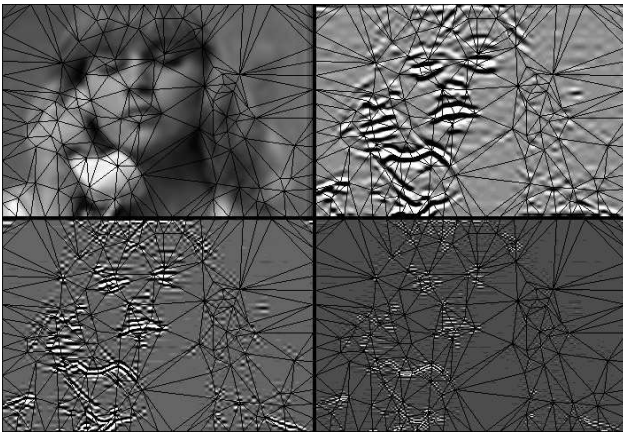


Figure 3: RDWT subbands and triangle mesh for the first frame of "Susie." Clockwise from upper-left: baseband; HL subband, scale 3; HL subband, scale 2; and HL subband, scale 1. A single triangle mesh is applied to all subbands at all orientations and scales, even though only the HL subbands are shown here.

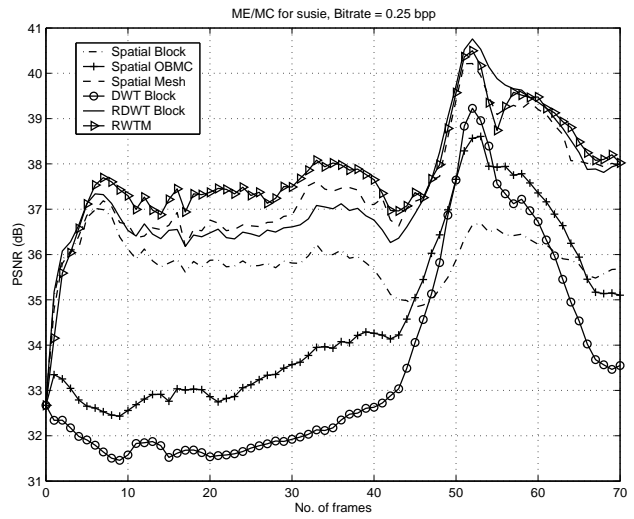


Figure 5: Frame-by-frame PSNR for "Susie" at 0.25 bpp.

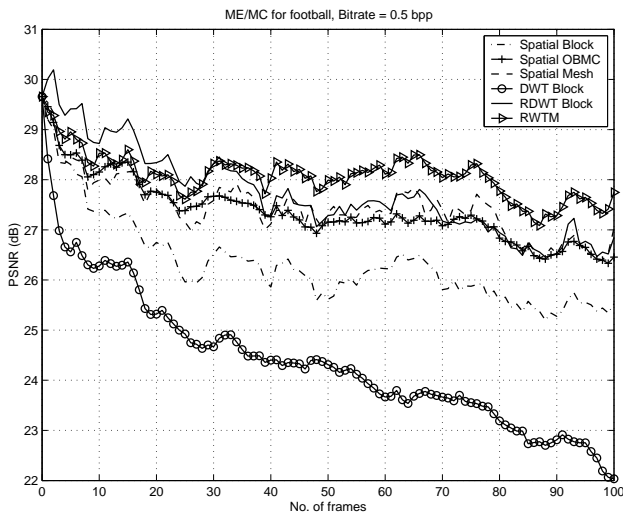


Figure 4: Frame-by-frame PSNR for "Football" at 0.5 bpp.