# MOTION COMPENSATION USING CRITICALLY SAMPLED DWT SUBBANDS FOR LOW-BITRATE VIDEO CODING

*Vildan Atalay Aydin, Hassan Foroosh*

University of Central Florida

## ABSTRACT

In this paper, we propose a novel motion estimation/motion compensation (ME/MC) method for wavelet-based (i.e. in-band) motion compensated temporal filtering (MCTF), with application to low-bitrate video coding. Unlike the conventional in-band MCTF algorithms, which require redundancy to overcome the shift-variance problem of critically sampled (i.e. complete) discrete wavelet transforms (DWT), we perform ME/MC steps directly on DWT coefficients by avoiding the need of shift-invariance. We omit upsampling, inverse-DWT (IDWT), and calculation of redundant DWT coefficients, while achieving arbitrary subpixel accuracy without interpolation, and high video quality even at very low-bitrates, by deriving the exact relationships between DWT subbands of input image sequences. Experimental results demonstrate the accuracy of the proposed method, confirming that our model for ME/MC effectively improves video coding quality.

***Index Terms***— Video Coding, Motion Compensated Temporal Filtering, Discrete Wavelet Transform

## 1. INTRODUCTION

Motion compensated temporal filtering is an essential step of numerous applications such as scalable video coding [1, 2], still image coding [3, 4], and denoising [5], to name a few. MCTF is performed either directly on input images, or on their transforms; thus, MCTF methods can be categorized into two groups depending on the order of temporal and spatial transforms.

MCTF techniques which perform temporal decomposition before a spatial transform include, Secker and Taubman [6], and Pesquest-Popescu and Bottreau [7] who use lifting formulation of three dimensional temporal wavelet decomposition for motion compensated video compression. Kim *et al.* [4] propose a 3-D extension of set partitioning in hierarchical trees (3D-SPIHT) for a low bit-rate embedded video coding scheme. More recently, Xiong *et al.* [8] extend spatiotemporal subband transform to in-scale motion compensation to exploit the temporal and cross-resolution correlations simultaneously, by predicting low-pass subbands from next lower resolution and high-pass subbands from neighboring frames in the same resolution layer. Furthermore, Chen and Liu [9] use an adaptive Lagrange multiplier selection model in rate-distortion optimization for motion estimation. In order to achieve more accurate motion data, Esche *et al.* [10] propose an interpolation method for motion information per pixel using block based motion data, and Rüfenacht *et al.* [11] anchor motion fields at reference frames instead of target frames to resolve folding ambiguities in the vicinity of motion discontinuities.

Although the performance of methods cited above are good, they have drifting and operational mismatch problems [8]. Therefore, performing spatial transform before temporal decomposition is introduced to overcome these drawbacks. However, since complete DWT is shift variant, in order to achieve in-band (i.e. directly in the wavelet domain) ME/MC, several methods are proposed to tackle this problem by redundancy. Van der Auwera *et al.* [1] use a bottom-up prediction algorithm for a bottom-up overcomplete discrete wavelet transform (ODWT). Park and Kim [2] propose a low-band-shift method by constructing a wavelet tree by shifting low-band subband in each level for horizontal, vertical, and diagonal directions for one pixel and performing downsampling. Andreopoulos *et al.* [3] define a complete to overcomplete discrete wavelet transform (CODWT), which avoids inverse DWT generally used to obtain ODWT. More recently, Liu and Ngan [12] use partial distortion search and anisotropic double cross search algorithms with the MCTF method of Andreopoulos *et al.* [3] for a fast motion estimation. Amiot *et al.* [5] perform MCTF for denoising, using dual-tree complex wavelet (DT-CW) coefficients.

All MCTF methods summarized above perform ME/MC either in the temporal domain before DWT, or in the wavelet domain with the help of redundancy (e.g. ODWT, DT-CW, etc.), due to the fact that complete DWT is shift-variant and motion estimation directly on DWT subbands is a challenging task. However, redundancy in these methods leads to high computational complexity [12]. Inspired by the fact that shift variance keeps the perfect reconstruction and nonredundancy properties of wavelets and breaks the coupling between spatial subbands, and that wavelet codecs always operate on complete DWT subbands [3], we propose a novel in-band ME/MC method, which avoids the need of shift invariance, and operates directly on the original DWT coefficients of the input sequences. Since Haar wavelets are widely utilized in MCTF

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

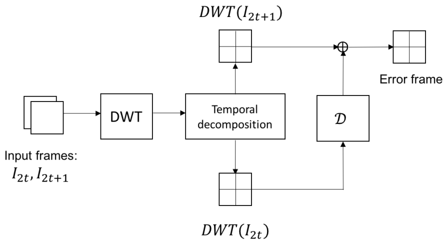

**Fig. 1**. Proposed in-band MCTF model.

methods due to the coding efficiency based on their short kernel filters [3], our method is built on Haar subbands. For accurate ME/MC, we define the exact relationships between the DWT subbands of input video sequences, which allows us to avoid upsampling, inverse DWT, redundancy, and interpolation for subpixel accuracy.

The rest of our paper is organized as follows. We introduce the problem and our proposed solution in Section 2. We define the derived exact inter-subband relationships in Section 3, demonstrate the experimental results in Section 4, and finally conclude our paper in Section 5.

## 2. MOTION COMPENSATED TEMPORAL FILTERING

In this section, we explain our proposed method for in-band MCTF, operating directly on DWT subbands.

The wavelet transform provides localization both in time and frequency; therefore, it is straightforward to use wavelets for MCTF. In order to perform ME/MC steps in MCTF, wavelet subbands of the transformed signal should be estimated. However, due to decimation and expansion operations of DWT, direct band-to-band estimation is generally not practical [2]. The proposed method overcomes this challenge by revealing the relationships between subbands of reference and target frames.

The proposed in-band MCTF method is demonstrated in Fig. 1. Given a video sequence, first, DWT is performed on each frame for spatial decomposition, then a temporal decomposition is performed by splitting video frames into groups. ME/MC ($\mathcal{D}$ in Fig. 1) is performed by block matching, using reference frames ($DWT(I_{2t})$) to predict the target frames ($DWT(I_{2t+1})$). Employing the found motion vectors (MV), reference frames are mapped onto the target frames to generate error frames as in;

$$E_{2t} = \text{DWT}(I_{2t+1}) - \mathcal{D}(\text{DWT}(I_{2t})) \tag{1}$$

where $E_{2t}$ stands for the error frame at time $2t$, which are then quantized and encoded/decoded by a wavelet codec, together with the MVs for the video coding application.

We employ Haar wavelet decomposition in spatial transform due to the benefits mentioned earlier. Since the method in Section 3 is accurate for any arbitrary subpixel translation defined as a multiple of $2^k$, where $k$ is the decomposition level, our method does not need interpolation for subpixel accuracy. A block matching method with unidirectional full search is used for ME/MC steps which is a common method used for MCTF. Our cost function is based on squared error minimization using all subbands, as follows:

$$\{\hat{x}, \hat{y}\} = \underset{x,y}{\arg\min}\{(A - \hat{A})^2 + (a - \hat{a})^2 + (b - \hat{b})^2 + (c - \hat{c})^2\}, \tag{2}$$

where, $\{\hat{x}, \hat{y}\}$ are the estimated motion vectors, $A, a, b, c$ denote the original target frame wavelet subbands, and $\hat{A}, \hat{a}, \hat{b}, \hat{c}$ are the predicted subbands for the same target image, using the method described in Section 3 and a reference frame.

## 3. INTER-SUBBAND RELATIONSHIP

In-band (i.e. wavelet domain) shift method along with the related notation are provided in this section.

### 3.1. Notation

Here, we provide the notations used throughout the paper, in Table 1.

**Table 1**. Notation

| | |
|---|---|
| $I_t$ | Input video frame at time $t$ |
| $A, a, b, c$ | Haar wavelet transform approximation, horizontal, vertical, and diagonal subbands of input image, respectively |
| $F, K, L$ | Matrices to be multiplied by approximation, horizontal, vertical, and diagonal DWT subbands, used for in-band shift of reference frame |
| $h$ | Number of hypothetically added levels in case of non-integer shifts |
| $s$ | Integer shift amount after the hypothetically added levels ($h$) |

Bold letters in the following sections demonstrate matrices and vectors. Subscripts $x$ and $y$ indicate the horizontal and vertical translation directions, respectively. Finally, subscript $k$ indicates the $k$th video frame, where $k = 1, 2, \ldots N$.

### 3.2. In-band Shifts

Our goal for the MCTF method described in Section 2 is to achieve ME/MC in the wavelet domain using DWT subbands, given a video frame sequence. For this purpose, wavelet subbands of the tranformed signal should be predicted using only DWT subbands of the reference frame. Therefore, we derive the relationship between the subbands of transformed and reference images, which can be described by in-band shift (i.e. in the wavelet domain) of the reference image subbands. Below, we derive the mathematical expressions which demonstrate these relationships [13].

Let $\mathbf{A}$, $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ be the first level approximation, horizontal, vertical, and diagonal detail coefficients (i.e. subbands), respectively, of a $2D$ reference frame at time $t$, $I_t$, of size $2m \times 2n$, where $m$ and $n$ are positive integers. Since decimation operator in wavelet transform reduces the size of input frame by half in each direction for each subband, we require the frame sizes to be divisible by 2. Now, the $1st$ level subbands of translated frame in any direction (i.e. horizontal, vertical, diagonal) can be expressed in matrix form using the $1st$ level Haar transform subbands of reference frame as in the following equations:

$$\begin{aligned}
\mathbf{A}_s &= \mathbf{F}_y\mathbf{A}\mathbf{F}_x + \mathbf{F}_y\mathbf{a}\mathbf{K}_1 + \mathbf{L}_1\mathbf{b}\mathbf{F}_x + \mathbf{L}_1\mathbf{c}\mathbf{K}_1 \\
\mathbf{a}_s &= -\mathbf{F}_y\mathbf{A}\mathbf{K}_1 + \mathbf{F}_y\mathbf{a}\mathbf{K}_2 - \mathbf{L}_1\mathbf{b}\mathbf{K}_1 + \mathbf{L}_1\mathbf{c}\mathbf{K}_2 \\
\mathbf{b}_s &= -\mathbf{L}_1\mathbf{A}\mathbf{F}_x - \mathbf{L}_1\mathbf{a}\mathbf{K}_1 + \mathbf{L}_2\mathbf{b}\mathbf{F}_x + \mathbf{L}_2\mathbf{c}\mathbf{K}_1 \\
\mathbf{c}_s &= \mathbf{L}_1\mathbf{A}\mathbf{K}_1 - \mathbf{L}_1\mathbf{a}\mathbf{K}_2 - \mathbf{L}_2\mathbf{b}\mathbf{K}_1 + \mathbf{L}_2\mathbf{c}\mathbf{K}_2
\end{aligned} \tag{3}$$

As already mentioned in Section 3.1, $\mathbf{F}$, $\mathbf{K}$, and $\mathbf{L}$ stand for matrices to be multiplied by the lowpass and highpass subbands of the reference frame in order to perform in-band shift, where subscripts $x$ and $y$ indicate *horizontal* and *vertical* shifts. $\mathbf{A}_s, \mathbf{a}_s, \mathbf{b}_s, \mathbf{c}_s$ are translated frame subbands (in any direction). The low/high-pass subbands of both reference and transformed frames are of size $m \times n$, $\mathbf{F}_y$ and $\mathbf{L}_{1,2}$ are $m \times m$, whereas $\mathbf{F}_x$ and $\mathbf{K}_{1,2}$ are $n \times n$.

By examining the translational shifts between subbands of two input frames in the Haar domain, we realize that horizontal translation reduces $\mathbf{L}$ to zero and $\mathbf{F}_y$ to the identity matrix. This could be understood by examining the coefficient matrices defined later in this section (namely, Eq. (4)), by setting the related vertical components to zero (specifically, $s_y$ and $h_y$). Likewise, vertical translation depends solely on approximation and vertical detail coefficients, in which case $\mathbf{K}$ is reduced to zero and $\mathbf{F}_x$ is equal to the identity matrix.

Here, we first define the matrices for subpixel shift amounts. The algorithm to reach any shift amount using the subpixel relationship will be described later in this section.

For subpixel translation, contrary to the customary model of approximating a subpixel shift by upsampling an image followed by an integer shift, our method models subpixel shift directly based on the original coefficients of the reference frame, without upsampling and the need for interpolation. To this end, we resort to the following observations:

**(1)** Upsampling an image $I$, is equivalent to adding levels to the bottom of its wavelet transform, and setting the detail coefficients to zero, while the approximation coefficients remain the same, as demonstrated in Fig. 2 for upsampling by $2^1$ as an example, where gray subbands show added zeros.

**(2)** Shifting the upsampled image by an amount of $s$ is equivalent to shifting the original image by an amount of $s/2^h$, where $h$ is the number of added levels (e.g. $h = 1$ in Fig. 2).

These observations allow us to do an in-band shift of the reference subbands for a subpixel amount, without upsampling or interpolation, which saves both memory and reduces the computational cost. Transformed signals therefore can be found by using the original level subbands of the reference image with the help of a hypothetically added level ($h$) and an integer shift value ($s$) at the added level.

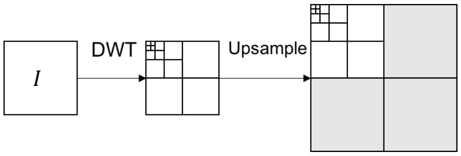

**Fig. 2**. Upsampling illustration.

Now, the aforementioned coefficient matrices, $\mathbf{F}_x$, $\mathbf{K}_1$, and $\mathbf{K}_2$ can be defined, in lower bidiagonal Toeplitz matrix form as follows.

$$\mathbf{F}_x = \frac{1}{2^{h_x+1}}\begin{bmatrix} 2^{h_x+1}-|s_x| & & & & \\ |s_x| & 2^{h_x+1}-|s_x| & & & \\ & |s_x| & & & \\ & & \ddots & \ddots & \\ & & & |s_x| & 2^{h_x+1}-|s_x| \end{bmatrix}$$

$$\mathbf{K}_1 = \frac{1}{2^{h_x+1}}\begin{bmatrix} -s_x & & & & \\ s_x & -s_x & & & \\ & s_x & & & \\ & & \ddots & \ddots & \\ & & & s_x & -s_x \end{bmatrix}$$

$$\mathbf{K}_2 = \frac{1}{2^{h_x+1}}\begin{bmatrix} 2^{h_x+1}-3|s_x| & & & & \\ -|s_x| & 2^{h_x+1}-3|s_x| & & & \\ & -|s_x| & & & \\ & & \ddots & \ddots & \\ & & & -|s_x| & 2^{h_x+1}-3|s_x| \end{bmatrix} \tag{4}$$

where $s_x$ and $h_x$ demonstrate the integer shift amounts at the hypothetically added level and the number of added levels for $x$ direction, respectively.

$\mathbf{F}_y$, $\mathbf{L}_1$, and $\mathbf{L}_2$ matrices are defined in a similar manner by upper bidiagonal Toeplitz matrices, using $y$ direction values for $s$ and $h$.

As mentioned earlier, $\mathbf{F}_x$ and $\mathbf{K}_{1,2}$ are $n \times n$, while $\mathbf{F}_y$ and $\mathbf{L}_{1,2}$ are $m \times m$. Sizes of these matrices also indicate that in-band shift of subbands is performed using only the original level Haar coefficients (which are of size $m \times n$) without upsampling. When the shift amount is negative, diagonals of the coefficient matrices interchange. For the MCTF method proposed in Section 2, the matrices are adapted for boundary condition by adding one more column/row at the end, where subband sizes are also adjusted to be $(m + 1) \times (n + 1)$.

The relationship defined above for subpixel shifts, can be used to produce any shift amount based on the fact that wavelet subbands are periodically shift-invariant. Table 2 demonstrates the calculation of any shift using subpixels, where $\%$ stands for modulo, $\lfloor s \rfloor$ and $\lceil s \rceil$ are the greatest integer lower than, and smallest integer higher than the shift amount $s$. Using circular shift of subbands for the given amounts in each shift amount case, and setting the new shift amount to the new shift values in Table 2, we can calculate any fractional or integer amount of shifts using subpixels.

**Table 2**. Arbitrary shifts defined by circular shift and subpixel amount

| Shift amount | Circular shift | New shift amount |
|---|---|---|
| $s\%2 = 0$ | $s/2$ | $0$ |
| $s\%2 = 1$ | $\lfloor s/2 \rfloor$ | $1$ |
| $\lceil s \rceil \%2 = 0$ | $\lceil s \rceil /2$ | $s - \lceil s \rceil$ |
| $\lfloor s \rfloor \%2 = 0$ | $\lfloor s \rfloor /2$ | $s - \lfloor s \rfloor$ |

If the shift amount (or the new shift amount in Table 2) is not divisible by 2, in order to reach an integer value at the $(N+h)$th level, the shift value at the original level is rounded to the closest decimal point which is divisible by $2^h$.

## 4. EXPERIMENTAL RESULTS

In this section, we demonstrate the results obtained with our method compared to the methods which perform in-band MCTF for video coding. We report our results on CIF video sequence examples with resolutions $352 \times 240$ and $352 \times 288$. We set our block size to $8 \times 8$ or $16 \times 16$ depending on the resolution of the sequences (in order to have integer number of blocks in subbands) and the required accuracy. Even though our MCTF method is based on 1-level DWT, we perform 2 more spatial decomposition levels after ME/MC steps before encoding, since compared methods use 3 spatial decomposition levels in total. Motion vectors and error frames are encoded using context-adaptive variable-length coding (CAVLC) and global thresholding with Huffman coding methods, respectively.

Fig. 3 shows the comparison of our method with respect to two conventional in-band methods, which are direct wavelet subband matching (i.e. band-to-band) and wavelet-block low-band-shift (LBS) [2] for CIF video sequence ”Football”. The graph demonstrates rate-distortion curves for a predicted frame of the Football sequence, where the shown bitrates are for error frame only (same as in the compared methods), and the accuracy for our method is set to $1/4$ pixel. As seen in this figure, our method improves PSNR compared to conventional in-band methods by $0.1 - 1$ dB.

We demonstrate our results for several video sequences at different bitrates in Fig. 4, where bitrates include the luminance component only, for the error frame, and MVs. The graph on the left shows the results with $1/2$ pixel accuracy using $16 \times 16$ blocks, and the one on the right uses $1/4$ pixel accuracy with $8 \times 8$ blocks. We also show the residual images for a predicted frame of the Foreman sequence in Fig. 5, for $0.1$ and $0.02$ bpp, respectively. The examples show how our method reduces the residual signal energy even at very low bitrates by providing more accurate reconstruction (prediction).

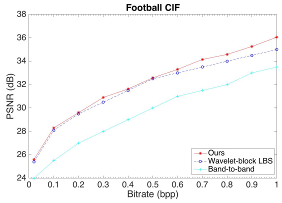

**Fig. 3**. Rate-distortion comparison for Football sequence.

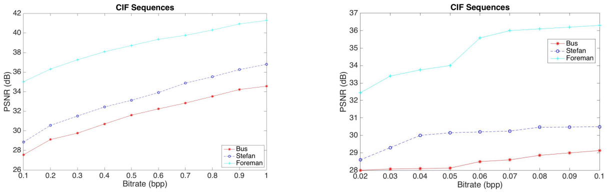

**Fig. 4**. PSNR performance of proposed method.

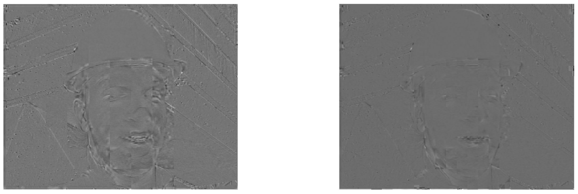

**Fig. 5**. Residual images for predicted frames of Foreman for $0.1$ bpp on the left and $0.02$ bpp on the right.

Finally, while the compared wavelet-block LBS method has 10 frames memory requirement for 3-level transform, our method requires only 1 frame memory (since upsampling is avoided). Computational complexity of our method is based on the matrix multiplications defined in Section 3, and full search method. We also reduce computational complexity by avoiding IDWT.

## 5. CONCLUSION

We propose a novel wavelet-based ME/MC method for MCTF focusing on low-bitrate video coding, where DWT is applied before temporal decomposition, and ME/MC steps are performed directly on DWT subbands. We avoid the need of shift-invariance property for non-redundant DWT (required by conventional methods), by deriving the exact relationships between DWT subbands of reference and transformed video frames. Our method avoids upsampling, inverse-DWT (IDWT), and calculation of redundant DWT while achieving high accuracy even at very low-bitrates. Experimental results show the accuracy of presented method, confirming that our model effectively improves video coding quality by reducing the residual energy in the error frames. The proposed ME/MC scheme can also be adapted for several image/video processing applications such as denoising, or scalable video coding.

## 6. REFERENCES

[1] G Van der Auwera, A Munteanu, P Schelkens, and J Cornelis, "Bottom-up motion compensated prediction in wavelet domain for spatially scalable video coding," *Electronics Letters*, vol. 38, no. 21, pp. 1251–1253, 2002.

[2] Hyun-Wook Park and Hyung-Sun Kim, "Motion estimation using low-band-shift method for wavelet-based moving-picture coding," *IEEE TIP*, vol. 9, no. 4, pp. 577–587, 2000.

[3] Yiannis Andreopoulos, Adrian Munteanu, Geert Van der Auwera, Jan PH Cornelis, and Peter Schelkens, "Complete-to-overcomplete discrete wavelet transforms: theory and applications," *IEEE Transactions on Signal Processing*, vol. 53, no. 4, pp. 1398–1412, 2005.

[4] Beong-Jo Kim, Zixiang Xiong, and William A Pearlman, "Low bit-rate scalable video coding with 3-d set partitioning in hierarchical trees (3-d spiht)," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 8, pp. 1374–1387, 2000.

[5] Carole Amiot, Catherine Girard, Jérémie Pescatore, Jocelyn Chanussot, and Michel Desvignes, "Fluoroscopic sequence denoising using a motion compensated multi-scale temporal filtering," in *ICIP*. IEEE, 2015, pp. 691–695.

[6] Andrew Secker and David Taubman, "Motion-compensated highly scalable video compression using an adaptive 3d wavelet transform based on lifting," in *ICIP*. IEEE, 2001, vol. 2, pp. 1029–1032.

[7] Béatrice Pesquet-Popescu and Vincent Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*. IEEE, 2001, vol. 3, pp. 1793–1796.

[8] Ruiqin Xiong, Jizheng Xu, and Feng Wu, "In-scale motion compensation for spatially scalable video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 2, pp. 145–158, 2008.

[9] Ying Chen and Guizhong Liu, "Adaptive lagrange multiplier selection model in rate distortion optimization for 3d wavelet-based scalable video coding," in *ICIP*. IEEE, 2014, pp. 3190–3194.

[10] Marko Esche, Michael Tok, and Thomas Sikora, "Adpative dense vector field interpolation for temporal filtering," in *ICIP*. IEEE, 2013, pp. 1918–1922.

[11] Dominic Rüfenacht, Reji Mathew, and David Taubman, "Hierarchical anchoring of motion fields for fully scalable video coding," in *ICIP*. IEEE, 2014, pp. 3180–3184.

[12] Yu Liu and King Ngi Ngan, "Fast multiresolution motion estimation algorithms for wavelet-based scalable video coding," *Signal Processing: Image Communication*, vol. 22, no. 5, pp. 448–465, 2007.

[13] Mais Alnasser and Hassan Foroosh, "Phase-shifting for nonseparable 2-d haar wavelets," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1061–1068, 2008.