AN EFFICIENT CONTENT-ADAPTIVE MC 3D-DWT WITH ENHANCED SPATIAL AND TEMPORAL SCALABILITY

Nagita Mehrseresht and David Taubman

The University of New South Wales, Sydney, Australia

ABSTRACT

In this paper we propose a novel, adaptive method for motion compensated 3D wavelet transformation (MC 3D-DWT) of video. The proposed method overcomes problems of ghosting and non-aligned aliasing artifacts which can arise in regions of motion model failure, when the video is reconstructed at reduced temporal or spatial resolutions. Previous MC 3D-DWT structures either take the form of a MC temporal DWT followed by a spatial transform ("t+2D"), or perform the spatial transform first, limiting the spatial frequencies which can be jointly compensated in the temporal transform, and hence limiting the compression efficiency. Essentially, the proposed transform continuously adapts itself between these two extremes, based on information available within the compressed bit-stream. Experimental results indicate that the proposed adaptive transform significantly reduces the cost in compression efficiency required to achieve high quality spatial and temporal scalability.

1. INTRODUCTION

Fully scalable video compression schemes use a feedforward compression paradigm in which a spatio-temporal (3D) discrete wavelet transform (DWT) is followed by embedded quantization and coding.

Without motion compensation, the order of the spatial and temporal DWT stages can be changed without altering the final subbands. However, it is not possible to efficiently exploit interframe redundancy without motion compensation. Unfortunately though, the commutative property of spatial (SDWT) and temporal (TDWT) DWT is lost with the introduction of motion compensation. Moreover, although advanced motion models are able to capture most of the actual scene activity, there are inevitably places (e.g., scene changes and occluded/uncovered regions) where the motion model must necessarily fail. When the motion model fails, the use of a low-pass temporal analysis filter results in ghosting artifacts under temporal scaling. Somewhat less obvious is the fact that under spatial scaling, reduced resolution sequences also suffer from motion-failure artifacts. Even motion model failure typically happens infrequently, these various artifacts can be quite annoying.

The problem of ghosting artifacts in the reduced frame-rate sequences has been addressed in [1][2]. While [1] proposes eliminating the low-pass filter, to eradicate the possibility of ghosting artifacts at reduced frame rates, the adaptive method proposed in [2] has significantly higher compression efficiency.

The problem of motion-failure artifacts in the reduced spatial resolution sequences has received relatively little attention. The problem can be avoided altogether by applying the motion compensated (MC) TDWT separately to the spatial subbands generated by an initial spatial decomposition of the original video frames. However, in [3] we find that structures of this form adversely affect the compression efficiency of the transform.

In this paper we propose a *fully content adaptive* method, which can remove motion-failure artifacts from both reduced spatial and reduced temporal resolution versions of a scalably compressed video sequence, while avoiding most of the efficiency losses associated with non-adaptive solutions. Essentially, the proposed transform adaptively adjusts both the temporal filters and the structure in which spatial and temporal transforms are combined. We build upon a flexible "2D+t+2D" structure for MC 3D-DWT, as proposed in [3]. This structure provides a framework in which we can selectively compensate for the effects of spatial aliasing. The adaptive scheme forms local estimates for the accuracy of the motion model, using this information to adjust a prescribed set of wavelet lifting steps. Although the estimates formed at the encoder and decoder may differ due to quantization effects and resolution scaling, these effects do not compromise the performance of the proposed scheme.

2. MOTION COMPENSATED TEMPORAL DWT

In this paper, we build upon the LIMAT [4] framework for MC TDWT. We use the deformable mesh motion model in connection with the bi-orthogonal 5/3 kernel due to its superior performance for temporal filtering. Using this framework, MC TDWT is accomplished through a sequence of temporal lifting steps and the motion compensation is performed *inside* each lifting step [4]. Let $\mathcal{W}_{k_1,k_2}(f_{k_1})$ denote a motion compensated mapping of frame f_{k_1} onto the coordinate system of frame f_{k_2}. Using this notation, we can implement the MC lifting steps for the 5/3 analysis by

$$h_k = f_{2k+1} - \frac{1}{2}[\mathcal{W}_{2k,2k+1}(f_{2k}) + \mathcal{W}_{2k+2,2k+1}(f_{2k+2})] \quad (1)$$

$$l_k = f_{2k} + \frac{1}{4}[\mathcal{W}_{2k-1,2k}(h_{k-1}) + \mathcal{W}_{2k+1,2k}(h_k)]. \quad (2)$$

Equations (1) and (2) are commonly known as "*prediction*" and "*update*" steps, respectively.

A reduced temporal resolution (frame rate) sequence may be formed by choosing the low-pass frames, l_k. High-pass frames h_k are the residual from bi-directional motion compensated prediction; their energy depends on the success of the motion model. In regions where the motion model captures the actual motion, the energy in the high-pass frames will be close to zero and the update steps will reduce the noise and aliasing in the low-pass frames. In regions where the motion model fails, however, the low pass temporal analysis filter is effectively being applied along invalid motion trajectories, resulting in visually annoying ghosting artifacts in the reduced frame-rate sequence. We can avoid these ghosting

0-7803-8554-3/04/$20.00 ©2004 IEEE.

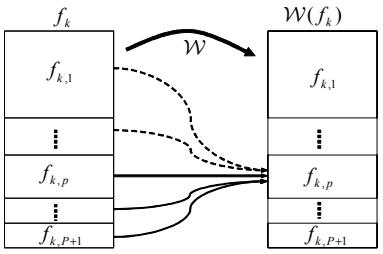

Fig. 1. Motion induced leakage between spatial resolution levels. Thin solid arrows: type I leakage. Dashed arrows: Type II leakage.

artifacts by skipping the update steps [2][1], so that

$$l_k = f_{2k} \tag{3}$$

but this has an inevitable adverse effect on the compression gain.

In [2] an adaptive temporal transform is proposed, in which different weights are assigned to the update steps based on the local estimated performance of the motion model (i.e., local energy in the high-pass temporal subbands). A weight of 0 results in equation (3), while the maximum weight reproduces equation (2).

Regardless of what motion model is used for $\mathcal{W}$, or how the lifting steps are weighted, the temporal transform can be trivially inverted by reversing the order of lifting steps and replacing summation with subtraction [4].

3. FLEXIBLE STRUCTURE FOR MC 3D-DWT

As mentioned before, changing the order of spatial and temporal decomposition stages in a MC 3D-DWT affects the generated spatio-temporal subbands. In the sequel, we consistently use the term "*resolution level*" (whether spatial or temporal) to refer to the collection of additional subband(s) required to double the resolution available from previous levels, where the lowest resolution level contains just the low-pass (spatial or temporal) subbands. We use bold face to refer to an entire sequence of frames from a resolution level. We use subscripts $_s$ to identify spatial resolution levels $\mathbf{f}_s$, and superscripts ^{t} to identify temporal resolution levels $\mathbf{f}^t$. Likewise, $\mathbf{f}_s^t$ denotes the spatio-temporal subbands produced after s levels of SDWT decomposition and t levels of MC-TDWT decomposition. $f_{k,s}^t$ refers to frame k in the sequence $\mathbf{f}_s^t$.

Shifting a frame even uniformly by a constant value σ, is not in general equal to separately shifting its subbands. Except for a uniform shift by an even number of pixels, shifting a frame changes the frequency content of its subbands. We refer to the motion induced leakage from lower resolution levels to higher resolution levels of the motion compensated frame as "*type I leakage.*" Similarly, "*type II leakage*" corresponds to the motion induced leakage from higher to lower resolution levels. These are illustrated in Fig. 1. The best prediction of a resolution level within the next frame, f_{k+1}, is found by motion compensating frame f_k and then finding the subbands of $\mathcal{W}_{k,k+1}(f_k)$. That is, we should, ideally, use the information from all subbands of f_k when predicting a subband of f_{k+1}. This is exactly what happens in the "t+2D" structure. Not surprisingly, then, if MC TDWT is applied separately within each resolution level, energy compaction and hence compression efficiency are reduced.

In this paper we build upon the "*flexible structure*" for the MC 3D-DWT proposed in [3]. This structure provides a framework in which we can choose the *order* and the *number* of spatial and MC temporal DWT decomposition stages. In this structure, P levels of SDWT decomposition are performed on the original video frames, yielding $P+1$ spatial resolution levels; T levels of MC TDWT are then applied to each spatial resolution level, followed by a further $S-P$ levels of SDWT, which are applied to the temporal subbands of the lowest initial spatial resolution level, $\mathbf{f}_{P+1}^t$. Using this structure, the original video sequence is eventually subjected to T levels of MC TDWT and S levels of SDWT. The value of P determines the number of levels of SDWT analysis which are performed prior to MC TDWT analysis. Choosing $P=0$ simplifies the transform to the "t+2D" structure, whereas $P=S$ yields a "2D+t" structure. In general, it is a *flexible "2D+t+2D"* structure.

For the lowest spatial resolution level $\mathbf{f}_{P+1}$, we can implement MC TDWT lifting steps in the same way as equations (1) and (2), except that we need to scale the motion parameters according to the spatial resolution of $\mathbf{f}_{P+1}$. For the remaining ($p \le P$) spatial resolution levels, $\mathbf{f}_p$ consists of the three high-pass subbands, HL_p, LH_p and HH_p. We cannot directly compensate for motion within these high-pass subbands, since shifting high-pass subbands does not produce the same linear phase relationship as shifting a baseband signal. We use the notation $\widetilde{\mathcal{W}}_{u,v}^p$ to refer to a modified motion compensation operator, which maps the three high-pass spatial subbands at spatial resolution level p of frame u onto the coordinate system of frame v. In the simplest case, we can implement $\widetilde{\mathcal{W}}_{u,v}^p(f_{u,p})$ using

$$\widetilde{\mathcal{W}}_{u,v}^p(f_{u,p}) = \mathcal{A}_H \circ \mathcal{W}_{u,v}^{p-1} \circ \mathcal{S}(0, f_{u,p}) \tag{4}$$

That is, we first synthesize the three subbands of frame $f_{u,p}$ into a baseband frame, using a single stage of spatial subband synthesis, denoted by the operator $\mathcal{S}$. For the moment, the missing LL subband required by $\mathcal{S}$ is treated as 0. Motion compensation is then applied to this synthesized frame, followed by spatial analysis back to the subband domain. The operator $\mathcal{A}_H$ denotes a single stage of spatial subband analysis, returning only the three high-pass subbands. The motion parameters associated with $\mathcal{W}_{u,v}^{p-1}$ are divided by 2^{p-1} to match the resolution of the synthesized baseband domain, within which motion compensation is performed. Neither "type I" nor "type II" leakage components from other spatial resolution levels have been used in equation (4). Using equation (4), the motion compensated lifting steps for resolution level f_p, $1 \le p \le P$, are given by

$$h_{k,p}^t = l_{2k+1,p}^{t-1} - \frac{1}{2}[\widetilde{\mathcal{W}}_{2k,2k+1}^p(l_{2k,p}^{t-1}) + \widetilde{\mathcal{W}}_{2k+2,2k+1}^p(l_{2k+2,p}^{t-1})] \tag{5}$$

$$l_{k,p}^t = l_{2k,p}^{t-1} + \frac{1}{4}[\widetilde{\mathcal{W}}_{2k-1,2k}^p(h_{k-1,p}^t) + \widetilde{\mathcal{W}}_{2k+1,2k}^p(h_{k,p}^t)] \tag{6}$$

Here, $\mathbf{l}_p^t$ and $\mathbf{h}_p^t$ denote the sequence of low-pass and high-pass *temporal subbands* produced by MC TDWT at temporal resolution t and spatial resolution level p, with $\mathbf{l}_p^0 \equiv \mathbf{f}_p$.

We cannot compensate for type II leakage effects without using information from higher frequency spatial subbands in the MC TDWT of $\mathbf{f}_p$. Such information is not available when inverting the temporal transform at a reduced spatial resolution. In fact, if type II leakage components are used in the motion compensation of some resolution level, these same components will appear when the video sequence is reconstructed at that resolution, since they cannot be compensated by the inverse temporal transform. This is effectively what happens in the "t+2D" structure. Subject to accurate motion modeling, it turns out that the type II leakage

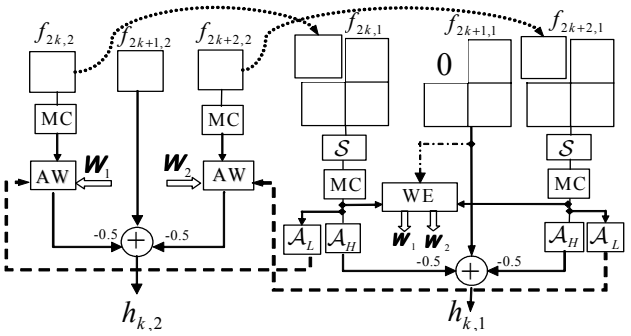

Fig. 2. Proposed adaptive prediction step.

components can even cancel some of the spatial aliasing which is produced naturally by the SDWT, reducing the total amount of aliasing power in the reduced spatial resolution sequence. When motion model fails, however, these aliasing components appear at misaligned spatial locations, producing the disturbing visual artifacts seen in Fig 5 (top).

We can, however, borrow information from lower spatial resolution levels to compensate for type I leakage, without affecting the appearance of reduced resolution frames. To compensate for type I leakage [3], we modify equation (4) to

$$\widetilde{\mathcal{W}}^{p}_{u,v}(l^t_{u,p}) = \mathcal{A}_H \circ \mathcal{W}^{p-1}_{u,v} \circ \mathcal{S}(l^t_u \widehat{LL}_p, l^t_{u,p}) \qquad (7)$$

Here, $l^t_u \widehat{LL}_p$ may be understood as an estimate of the missing low-pass subband corresponding to the high-pass spatial subbands in $l^t_{u,p}$. Although compensating for type I leakage as in equation (7) significantly improves the compression efficiency, we still lose substantial amount of compression gain due to the uncompensated type II leakage components [3].

4. ADAPTIVELY TYPE II LEAKAGE COMPENSATION

In this Section, we propose a novel adaptive scheme which resolves the problem of non-aligned spatial aliasing artifacts and almost preserves the compression efficiency of the "t+2D" structure. We adaptively compensate for type II leakage based on the *estimated local* performance of the motion model. Fig. 2 shows the adaptive prediction lifting steps for $P = 1$. Dotted arrows show the information borrowed to compensate for type I leakage in the manner of equation (7). Dashed arrows illustrate the interaction required for adaptive type II leakage compensation. The operator $\mathcal{A}_L$ refers to a single stage of SDWT decomposition, returning the LL subband.

At each spatial location, we find a corresponding weight based on a local estimate of the motion accuracy. To avoid explicitly sending the weights, the weight estimation (WE) method must be reasonably robust to quantization error. Also, at full resolution when there is no quantization error, the decoder must be able to recover exactly the same weights. For WE, we use a similar method to that proposed for update steps in [2]. WE can be based on a single motion compensation operator, or a pair of motion compensation operators (forward and backward). In the first case, we can use the local energy in the motion prediction residual, $\mathcal{A}_H \circ \mathcal{W}^{p-1}_{2k,2k+1} \circ \mathcal{S}(0, f_{2k,p}) - f_{2k+1,p}$. This scheme is suitable for any lifting-based temporal transform, including the Haar.

Unfortunately, when WE is based only on a single operator $\mathcal{W}$, we cannot use lower frequency information to improve the robustness of the estimator. We can avoid this problem by basing the WE on a pair of motion compensation operators. This is appropriate where bi-directional motion compensation is employed, as with the $5/3$ transform. In this case, the weights are found based on the difference between the MC versions of two successive even indexed frames, $\mathcal{W}^{p-1}_{2k,2k+1} \circ \mathcal{S}(f_{2k}\widehat{LL}_p, f^t_{2k,p}) - \mathcal{W}^{p-1}_{2k+2,2k+1} \circ \mathcal{S}(f_{2k+2}\widehat{LL}_p, f^t_{2k+2,p})$. In this way, we fully exploit the available lower resolution information which is already being borrowed for type I leakage compensation, as illustrated in Fig. 2. To further reduce the sensitivity to quantization error, we use a spatial averaging window. We also use a mapping function (e.g., a non-linear quantizer) to map the average local energy to the desired weight.

After estimating the weights, the adaptive weighting (AW) in Fig. 2 is implemented as

$$\begin{aligned} h^t_{k,2} \;=\; & l^{t-1}_{2k+1,2} - \frac{1}{2}[(1 - \mathbf{W}_1) \times \mathcal{W}^1_{2k,2k+1}(l^{t-1}_{2k,2}) + \\ & \mathbf{W}_1 \times \mathcal{A}_L \circ \mathcal{W}_{2k,2k+1} \circ \mathcal{S}(l^{t-1}_{2k}\widehat{LL}_1, l^{t-1}_{2k,1})] + ... \end{aligned} \qquad (8)$$

In equation (8) we write "..." to represent the similar contribution from the backward motion compensation (i.e., frame $l^{t-1}_{2k+2,1}$). Also we have used normalized weights in the range 0 to 1.

When the motion model captures the scene activity, the weights are equal to 1 so that the prediction step of equation (8) takes advantage of the information from higher frequency spatial subbands (type II leakage components). On the other hand, when the motion model fails, the adaptive algorithm assigns a weight of 0 to the type II leakage components. In this case the adaptive prediction steps in Fig. 2 exclude all type II leakage information, thereby avoiding the possibility of non-aligned aliasing artifacts at reduced spatial resolutions. In practice, behaviour between these two extremes is addressed by assigning various weights in the range 0 to 1, based on the likelihood of motion failure.

It is sufficient to adaptively compensate for type II leakage *just* within the *prediction steps*. The adaptive leakage compensation scheme is beneficial at locations where motion modeling is successful. In this case, the total energy and consequently the inter-resolution leakage in the high-pass temporal frames are small and we do not need to compensate for inter-resolution leakage. Although Fig. 2 and equation (8) corresponds to $P = 1$, the extension to $P > 1$ is easy to achieve.

5. ADAPTIVELY WEIGHTING THE UPDATE STEPS

To remove ghosting artifacts from reduced frame-rate sequences, we extend the adaptive method of [2] to the "2D+t+2D" structure. Fig. 3 illustrates the required extension when $P = 1$. Again, extending to more than two spatial resolution levels ($P > 1$) is trivial, as the update steps in each spatial resolution level are independent of other resolution levels. This extension also addresses the problem of adaptive temporal update weighting in the presence of spatial scalability, which was not addressed in [2].

6. EXPERIMENTAL RESULTS

In this Section, we experimentally investigate the compression efficiency and the spatio-temporal scalability of the proposed adaptive scheme. For an SNR (quality/bit-rate) scalable video compres-

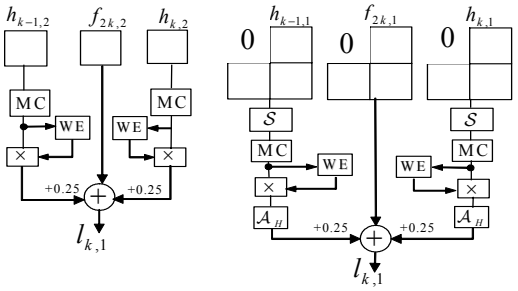

Fig. 3. Extending the adaptive update weighting scheme to each spatial resolution level.

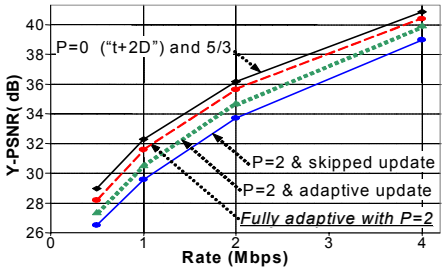

Fig. 4. Luminance PSNR of Stefan sequence at different bit-rates.

sion scheme, the spatio-temporal subbands must further be subjected to embedded quantization and coding. We use the EBCOT algorithm [5] embodied within the JPEG2000 image compression standard to create an efficient embedded codestream for spatio-temporal subbands.

Fig. 4 illustrates the luminance (Y) PSNR of the reconstructed full resolution (CIF, 30fps) standard test video sequence "Stefan" when scalably reconstructed at different bit-rates. We use $T = 3$ levels of MC TDWT and $S = 5$ levels of SDWT. As we expect, the "t+2D" structure ($P = 0$) with MC $5/3$ TDWT has the highest compression efficiency. Fig. 5 (top), however, illustrates a sample frame from reconstructed QCIF resolution full frame-rate sequence when we use "t+2D" structure without quantizing the spatio-temporal subbands. The non-aligned spatial aliasing artifacts are obvious specially where motion model fails.We can avoid these artifacts by using pre-temporal spatial decomposition ($P > 0$) in the "2D+t+2D" structure. However, the PSNR continues to drop significantly by increasing P. Also to avoid the ghosting artifacts in the reduced frame-rate sequences, we skip the update steps (equation (3)) as it is shown in Fig. 4. Using the proposed fully adaptive scheme significantly improves the compression efficiency of the transform. In addition, as Fig. 5 (down) shows, by using the proposed fully adaptive transform there is no visually annoying misaligned artifacts in the reconstructed reduced spatial resolution sequence. Visually annoying ghosting artifacts in the reduced frame-rate sequences are also successfully removed by extending the adaptive update weighting scheme of [2] to the flexible structure with $P > 0$. Since the space is limited, we omit the illustration of this previously well-known capability of the temporal adaptive schemes. In general, by using the adaptive transform achieving high quality spatial and temporal resolution scalability has much smaller cost on the compression efficiency.

Fig. 5. Sample reconstructed frame at QCIF resolution using the "t+2D" structure (top) and adaptive transform (down).

7. CONCLUSIONS

In this paper, we have significantly extended the ideas in [2] to demonstrate a fully scalable content adaptive MC 3D-DWT, which achieves high compression efficiency while removing the disturbing motion-failure artifacts which can arise when the video is reconstructed at reduced spatial and/or temporal resolutions. In our opinion, the proposed adaptive structure resolves a key problem in the development of highly efficient fully scalable video compression systems.

8. REFERENCES

[1] M. V. der Schaar and D. Turaga, "Unconstrained motion compensated temporal filtering (UMCTF) framework for wavelet video coding," *IEEE Int. Conf. Accoust. Speech and Signal Proc.*, pp. 81–84, 2003.

[2] N. Mehrseresht and D. Taubman, "Adaptively weighted update steps in motion compesated lifting based scalable video compression," *IEEE Int. Conf. Image Proc.*, pp. 771–774, 2003.

[3] ——, "Spatial scalability and compression efficiency within a flexible motion compensated 3d-DWT," *Accepted for IEEE Int. Conf. Image Proc.*, 2004.

[4] A. Secker and D. Taubman, "Lifting based invertible motion adaptive transform, LIMAT, framework for highly scalable video compression," *IEEE Trans. Image Proc.*, vol. 12, pp. 1530–1542, Dec. 2003.

[5] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Trans. Image Proc.*, vol. 9, no. 7, pp. 3445–3462, July 2000.