SPATIAL SCALABILITY AND COMPRESSION EFFICIENCY WITHIN A FLEXIBLE MOTION COMPENSATED 3D-DWT

Nagita Mehrseresht and David Taubman

The University of New South Wales, Sydney, Australia

ABSTRACT

We investigate the implications of the conventional "t+2D" MC 3D-DWT structure for spatial scalability, and propose a more flexible "2D+t+2D" structure. An initial P levels of spatial wavelet decomposition are followed by T levels of motion compensated temporal decomposition, applied separately to each spatial resolution level. A further $S - P$ levels of spatial decomposition are applied to the resulting subbands. By adjusting P, the structure allows us to trade energy compaction with the potential for artifacts at reduced spatial resolutions. This allows us to study the interaction between scalability and compression efficiency. We show that the "t+2D" structure ($P = 0$) necessarily maximizes compression efficiency, while allowing misaligned spatial aliasing artifacts to arise at reduced resolutions. These artifacts can be removed by increasing the value of P, at an inevitable cost in compression efficiency. We show how this cost can be minimized.

1. INTRODUCTION

Recently a large amount of effort has been invested in wavelet-based fully scalable video coding. Interactive multimedia, video surveillance and video delivery over heterogeneous networks and/or in error prone environments can be named as some of the applications which stand to benefit from this research. The predictive feedback paradigm inherent in traditional video compression algorithms is incompatible with the requirement of highly scalable compression. Instead, the preferable paradigm is that of feed-forward compression in which a spatio-temporal (3D) discrete wavelet transform (DWT) is followed by embedded quantization and coding. To exploit interframe redundancy, the transform must compensate for the motion between frames.

Without motion compensation, the 3D-DWT can be formed by separable extension of the 1D-DWT to the spatial and temporal dimensions. In this case, spatial and temporal DWT stages can be performed in any desired order, without altering the final subbands. With the introduction of motion compensation, however, this commutative property is lost, principally because the spatial wavelet transform is not shift invariant. By changing the order of the spatial and motion compensated temporal DWT (MC TDWT) operators, a family of different MC 3D-DWTs may be formed.

Early work on scalable MC 3D-DWT followed the so-called "*t+2D*" paradigm, in which MC TDWT is applied to the full resolution video frames, followed by a spatial DWT (SDWT) of the temporal subbands [1][2]. New schemes proposed in [3][4] also fall into this category. In [5][6] an alternate, "*2D+t*" approach may be found, in which MC TDWT is applied to the spatial subbands generated by spatial decomposition of the original video frames.

Until this time, very little effort has been invested in directly comparing different structures for MC 3D-DWT. In this paper we propose a flexible "*2D+t+2D*" structure for MC 3D-DWT, in which any desired number of spatial decomposition levels are performed on the original full resolution video, followed by MC TDWT of the spatial subbands. Further levels of spatial decomposition may be performed on the resulting subbands to provide additional levels of spatial scalability and energy compaction. Using this structure, we can flexibly choose the particular MC 3D-DWT which is most suited to the desired application. Perhaps more significant than the structure itself is the insight which it provides into the interaction between spatial aliasing, scalability and energy compaction. We first suggested the "*2D+t+2D*" structure in [7], along with some preliminary results. The present paper is concerned with a more careful investigation into the interaction between spatial scalability and compression efficiency.

2. MOTION COMPENSATED TEMPORAL LIFTING

We build upon the LIMAT [4] framework for MC TDWT, using the deformable mesh motion model in connection with the bi-orthogonal 5/3 kernel, due to its superior performance for temporal filtering. However, our approach is suitable for other motion models and wavelet kernels. Using this framework, MC TDWT is accomplished through a sequence of temporal lifting steps. To exploit inter-frame dependence, we compensate for motion *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 filters 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.

Regardless of what motion model and interpolation operators are used for $\mathcal{W}$, the temporal transform can be trivially inverted by reversing the order of the lifting steps and replacing addition with subtraction, as indicated by equations (3) and (4)

$$f_{2k} = l_k - \frac{1}{4}[\mathcal{W}_{2k-1,2k}(h_{k-1}) + \mathcal{W}_{2k+1,2k}(h_k)] \quad (3)$$

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

3. FLEXIBLE "2D+T+2D" STRUCTURE

The purpose of MC 3D-DWT analysis is to generate a spatio-temporal multi-resolution representation for the original video sequence. In the sequel, we consistently use the term "*resolution*

0-7803-8554-3/04/$20.00 ©2004 IEEE.

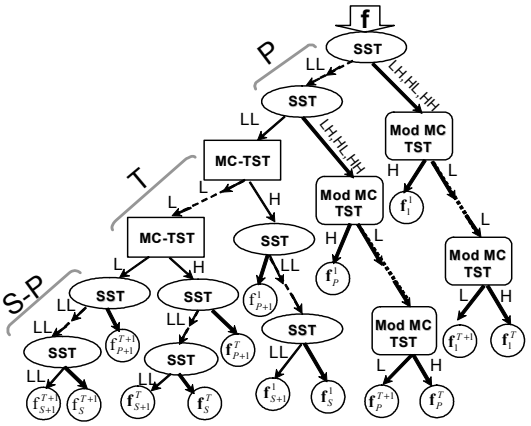

Fig. 1. Proposed scheme for MC 3D-DWT analysis. SST and TST stands for spatial and temporal subband transforms, respectively.

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. We also refer to frame k in the sequence $\mathbf{f}_s^t$ by $f_{k,s}^t$.

In the proposed structure, we apply MC TDWT *separately* to each spatial resolution level produced by an initial spatial decomposition (SDWT) of the original video frames. Fig. 1 indicates the analysis part of the proposed scheme. In this structure, P levels of SDWT are performed on the original video frames, yielding $P+1$ spatial resolution levels; T levels of MC TDWT are then applied separately 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 spatial analysis which are performed prior to MC temporal analysis. Choosing $P=0$ simplifies the transform to the "t+2D" structure, whereas $P=S$ yields the "2D+t" structure.

In this paper, we are interested in the effect of different structures on scalability and energy compaction; therefore, we use same set of motion parameters, regardless of the value of P. We note, however, that it is possible to employ $P+1$ different sets of motion parameters, since MC TDWT is applied separately to each of spatial resolution levels $\mathbf{f}_1$ through $\mathbf{f}_{P+1}$. This property can potentially be exploited for a spatially scalable motion representation.

The MC TDWT of each initial spatial resolution level $\mathbf{f}_p$ deserves further explanation. 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

$$\widetilde{\mathcal{W}}_{u,v}^p(f_{u,p}) = \mathcal{A}_H \circ \mathcal{W}_{u,v}^{p-1} \circ \mathcal{S}(0, f_{u,p}) \qquad (5)$$

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.

Using equation (5), the motion compensated lifting steps for resolution level $\mathbf{f}_p$, $1 \le p \le P$, are given by

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

$$l_{k,p}^t = l_{2k,p}^{t-1} + \frac{1}{4}[\widetilde{\mathcal{W}}_{2k-1,2k}^p(h_{k-1,p}^t) + \widetilde{\mathcal{W}}_{2k+1,2k}^p(h_{k,p}^t)] \qquad (7)$$

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 $\mathbf{l}_p^0 \equiv \mathbf{f}_p$. While the above scheme for MC TDWT of spatial resolution levels is independent of the motion model, we note that a particularly efficient implementation structure has been proposed in [6], for the special case of block-based motion.

4. SPATIO-TEMPORAL SCALABILITY

We can reconstruct the video signal to any desired spatial and temporal resolution, by reversing the processing flow of the analysis transform and replacing the spatial and temporal analysis operations with their corresponding synthesis operations.

With $P=0$ ("t+2D" structure), temporal scalability is easily achieved by terminating MC TDWT synthesis at the desired temporal resolution. When $P>0$, we can achieve temporal scalability by terminating all $P+1$ MC TDWT synthesis operations at the desired temporal resolution. It should be noted that when $P>0$, different spatial resolution levels of the original video sequence have been separately subjected to MC TDWT. This causes the reduced temporal resolution sequence to be different when different values of P are used. In fact, temporal scalability with $P>0$ can introduce undesirable spatial aliasing effects around strong image edges, in addition to the ghosting artifacts which are sometimes observed with the "t+2D" structure in regions of motion model failure. It can be shown, however, that both types of artifact arise only from the update lifting steps. By skipping the update steps or (preferably) using an adaptive update weighting scheme, such as that proposed in [8], it is possible to produce high quality temporally scaled sequences without sacrificing compression efficiency, even when $P>0$.

When $P=0$ ("t+2D" structure), spatial scalability can be achieved by terminating the spatial synthesis process at the desired resolution and applying MC temporal synthesis to the reduced spatial resolution frames, after appropriate scaling of the motion parameters. When $P>0$, spatial scalability can be achieved in the

first instance by discarding up to the first P spatial resolution levels. Lower spatial resolutions may be obtained (when $S > P$), in a similar manner to the way in which the spatial scalability is achieved with the "t+2D" structure. Thus, for resolution reduction factors up to 2^P, the low resolution video sequences are identical to those obtained by extracting reduced spatial resolutions from the original video frames. For larger reduction factors, however, *non-aligned spatial aliasing artifacts* can appear within the spatially scaled video, in regions where the motion model fails.

To understand the origin and nature of these non-aligned spatial aliasing artifacts, consider the case $P = 0$ and suppose we reconstruct the video at half its original spatial resolution, writing f'_k for the resulting video frames. Ignoring the typically small differences between f'_{2k} and $\mathcal{A}_L(f_{2k})$ (there is no difference if update steps are skipped), we obtain

$$\begin{aligned} f'_{2k+1} &\approx \mathcal{A}_L(h_k) + \frac{1}{2}\mathcal{W}^1_{2k,2k+1} \circ \mathcal{A}_L(f_{2k}) + ... \\ &\approx \mathcal{A}_L(h_k) + \frac{1}{2}\mathcal{A}_L \circ \mathcal{W}_{2k,2k+1} \circ \mathcal{S}(\mathcal{A}_L(f_{2k}), 0) + ... \\ &= \mathcal{A}_L(f_{2k+1}) - \frac{1}{2}\mathcal{A}_L \circ \mathcal{W}_{2k,2k+1}[f_{2k} - \mathcal{S}(\mathcal{A}_L(f_{2k}), 0)] - ... \\ &= \mathcal{A}_L(f_{2k+1}) - \frac{1}{2}\mathcal{A}_L \circ \mathcal{W}_{2k,2k+1} \circ \mathcal{S}(0, \mathcal{A}_H(f_{2k})) - ... \end{aligned}$$

In the above, we use $\mathcal{A}_L$ to denote one level of spatial wavelet analysis, retaining only the LL subband, and we write "..." to represent contributions from f_{2k+2}; these are analogous to those from f_{2k}. Evidently, the difference between f'_{2k+1} and $\mathcal{A}_L(f_{2k+1})$ arises from the spatial aliasing components which remain after synthesizing the high-pass subbands of f_{2k} (resp. f_{2k+2}), motion compensating them, and taking their low-pass subbands.

The appearance of spatial aliasing terms is not surprising. Any form of spatial resolution reduction, including that associated with $\mathcal{A}_L(f_{2k+1})$, incurs some aliasing. In fact, subject to accurate motion modeling, it can be shown [9] that f'_{2k+1}typically has even less total aliasing power than $\mathcal{A}_L(f_{2k+1})$! Wherever the motion model fails, however, the aliasing terms produced by $\mathcal{A}_L \circ \mathcal{W}_{2k,2k+1} \circ \mathcal{S}(0, \mathcal{A}_H(f_{2k}))$ do not line up with the scene features (typically edges) which generated them. The resulting, non-aligned aliasing artifacts, can be visually disturbing.

5. LEAKAGE COMPENSATION AND EFFICIENCY

In the special case of ideal bandlimited spatial subband filters, and pure translational motion, the synthesized baseband signal $\mathcal{S}(0, f_{u,p})$ in equation (5) is perfectly bandlimited. This, coupled with the fact that $\mathcal{W}^{p-1}_{u,v}$ is an LTI operator, means that no information is lost when $\mathcal{A}_H$ is applied to $\mathcal{W}^{p-1}_{u,v} \circ \mathcal{S}(0, f_{u,p})$. In this special case, it is not hard to show [9] that the temporal subbands produced using equations (6) and (7) are identical to those produced by the "t+2D" structure.

In practice, the subband filters are not perfectly bandlimited and real motion can be locally expansive or contractive. These factors generally result in the loss (or leakage) of information during the implementation of $\widetilde{\mathcal{W}}^p_{u,v}$. Of course, this does not affect the invertibility of the complete transform, which is trivially obtained by replacing $\mathcal{W}_{u,v}$ with $\widetilde{\mathcal{W}}^p_{u,v}$ in equations (3) and (4). It does, however, reduce the energy compaction and hence compression efficiency associated with the motion compensated temporal transform.

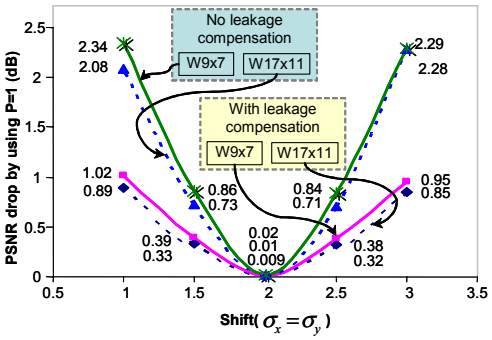

Fig. 2. Loss in compression performance when $P = 1$ relative to $P = 0$, using a synthetic video sequence with ideal motion.

To see this, consider the prediction step in equation (6), with $t = 1$ so that $\mathbf{l}^{t-1}_p = \mathbf{f}_p$. Assuming a perfect motion model, $f_{2k+1,p}$ would be pefectly compensated if we could motion compensate the original high resolution frames f_{2k} and f_{2k+2}, finding spatial resolution p of the result. Instead of motion compensating the original frames, however, we are only using the information contained in $f_{2k,p}$ and $f_{2k+2,p}$. In so doing, we are unable to exploit the motion induced leakage of information from lower frequency subbands of $\mathbf{f}$ to $\mathbf{f}_p$ (call this "*type I leakage*"), and we are unable to exploit the motion induced leakage of information from higher frequency subbands of $\mathbf{f}$ to $\mathbf{f}_p$ (call this "*type II leakage*"). As a result, the energy of the high-pass temporal subbands $\mathbf{h}^t_p$ is generally higher than in the case of the "t+2D" structure, leading to reduced compression efficiency.

Fig. 2 provides an experimental demonstration of these effects, in the context of a synthetic video sequence created by cropping a sequence of shifted CIF resolution windows from a much larger image. The horizontal axis in the figure represents the amount of shift between successive frames. The vertical axis measures the drop in compression efficiency (measured in PSNR) of a complete compression system (see Section 6) when going from $P = 0$ to $P = 1$ level of pre-temporal spatial decomposition. The PSNR values correspond to 1Mbps compressed bit-rate. As expected, there are no leakage effects when the shift is an even number of pixels, since the spatial subband transform is periodically shift invariant with period 2. Leakage effects are greatest when the shift is an odd number of pixels. Also as expected, the use of longer subband filters (17x11 vs. 9x7), with sharper frequency responses, can reduce the aliasing leakage.

The lower curves in Fig. 2 are obtained by compensating for type I leakage effects. 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. We can, however, borrow information from lower spatial resolution levels to compensate for type I leakage. This does not sacrifice the desirable property that spatial scalability, with reduction factors up to 2^P, produces video sequences identical to those obtained by reducing the spatial resolution of the original video frames.

To compensate for type I leakage, we modify equation (5) to

$$\widetilde{\mathcal{W}}^p_{u,v}(l^t_{u,p}) = \mathcal{A}_H \circ \mathcal{W}^{p-1}_{u,v} \circ \mathcal{S}(\widehat{l^t_u LL_p}, l^t_{u,p}) \qquad (8)$$

Here, $\widehat{l^t_u LL_p}$ may be understood as an estimate of the missing low-

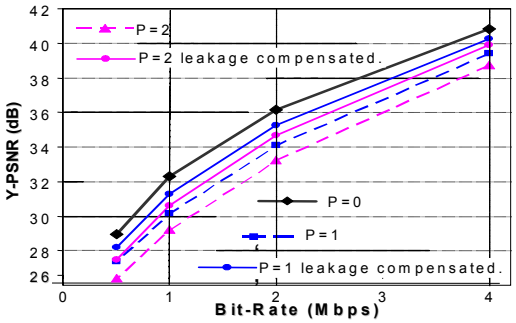

Fig. 3. Luminance PSNR of the "Stefan" sequence at full resolution using $T = 3$, $S = 5$ and $P = \{0, 1, 2\}$.

pass subband. The idea is to emulate the behaviour of applying the baseband MC TDWT directly to the LL video sequence produced by a $(p-1)$-level SDWT, followed by application of the $\mathcal{A}_H$ operator. It is possible to do this perfectly for $t = 1$. More generally, however, we must restrict our estimate of $l_u^t \widehat{LL}_p$ to use information only from the corresponding temporal resolution, at lower spatial resolutions. We do this by synthesizing the lower resolution subbands available from corresponding lifting steps at the same temporal resolution, i.e.,

$$l_u^t \widehat{LL}_p = \begin{cases} l^t_{u,P+1} & p = P \\ \mathcal{S}(l^t_{u,P}, l^t_{u,P+1}) & p = P-1 \geq 1 \\ \mathcal{S}(l^t_{u,P-1}, \mathcal{S}(l^t_{u,P}, l^t_{u,P+1})) & p = P-2 \geq 1 \\ \ldots & \ldots \end{cases}$$

6. EXPERIMENTS WITH REAL VIDEO SEQUENCES

In this section, we compare the compression efficiency of the MC 3D-DWT when different values of P are used. For an SNR (quality/bit-rate) scalable video compression scheme, the spatio-temporal subbands must further be subjected to embedded quantization and coding. This allows the quality of the overall video sequence to be adjusted at each spatio-temporal resolution by discarding appropriate subsets from the embedded bitstream. We use the EBCOT algorithm [10] embodied within the JPEG2000 image compression standard to create an efficient embedded codestream for spatio-temporal subbands.

The standard CIF resolution test sequence "Stefan" is compressed using MC 3D-DWT with $T = 3$ levels of MC TDWT and $S = 5$ levels of 9/7 SDWT. Fig. 3 illustrates the luminance (Y) PSNR of the reconstructed full resolution sequence at different bit-rates, when different numbers of pre-temporal spatial decomposition levels P are used. This figure confirms the results obtained in Fig. 2, although that figure was obtained with a simulated source and ideal motion, having abnormally high energy compaction which tends to exaggerate relative losses due to leakage. As before, we see that $P = 0$ ("t+2D" structure) yields the highest compression efficiency. Moreover, the efficiency continues to drop as P is increased. Fig. 3 also confirms the benefits of leakage compensation.

Our choice of the "Stefan" sequence here is motivated by the fact that it exhibits complex motion, with some motion model failure. In this case, $P > 0$ is required if non-aligned aliasing artifacts are to be avoided at reduced spatial resolutions.

7. CONCLUSIONS

In this paper we proposed a flexible structure for MC 3D-DWT, which enables us to trade energy compaction with the potential for artifacts at reduced spatial resolutions by choosing the number of pre-temporal spatial decomposition levels P. We found that compression performance of the MC 3D-DWT is maximized by using the "t+2D" structure ($P = 0$). On the other hand, we also showed that such a structure can lead to visually annoying artifacts, in places where the motion model fails. Using the flexible structure, we can avoid these artifacts by selecting an appropriate value for P. We found that the performance loss associated with $P > 0$ is mainly caused by spatial frequency leakage during motion compensation. We showed that this phenomenon can be reduced by leakage compensation and, to a certain extent, by the selection of subband filters with better frequency roll-off performance. Since space is limited, in this paper we report results only for the "Stefan" sequence; however, we have tested the proposed structure with a wide range of synthetic and standard CIF and SIF resolution video sequences.

8. REFERENCES

[1] K. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Trans. Image Proc.*, vol. 3, no. 5, pp. 559–571, Sept. 1994.

[2] S. Choi and J. Woods, "Motion compensated 3d subband coding of video," *IEEE Trans. Image Proc.*, vol. 8, pp. 155–167, Feb 1999.

[3] V. Bottreau, M. Benetiere, B. Felts, and B. Pesquet-Popescu, "A fully SCalable 3d subband video codec," *IEEE Int. Conf. Image Proc.*, pp. 1017–1020, 2001.

[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] J. Woods and G. Linlienfield, "A resolution and frame-rate scalable Subband/Wavelet video coder," *IEEE Tran. Circ. and Syst. for Video Tech.*, vol. 11, no. 9, pp. 1035–1044, Sept 2001.

[6] Y. Andreopoulos, M. Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and J. Cornelis, "Complete-to-overcomplete discrete wavelet transforms for scalable video coding with MCTF," *SPIE Visual Comm. Image Proc.*, pp. 719–731, 2003.

[7] D. Taubman, "Succesive refinement of video: Fundamental issues, past efforts and new directions," *SPIE Visual Comm. and Image Prog.*, pp. 649–663, 2003.

[8] 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.

[9] ——, "A flexible structure for fully scalable motion compensated 3d-DWT with emphasis on the impact of spatial scalability," *submitted to IEEE Tran. Image Proc.*, 2004.

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