UNCONSTRAINED MOTION COMPENSATED TEMPORAL FILTERING (UMCTF) FRAMEWORK FOR WAVELET VIDEO CODING

M. van der Schaar and D. S. Turaga
Philips Research USA
{mihaela.vanderschaar, deepak.turaga}@philips.com

ABSTRACT

This paper presents a new framework for adaptive temporal filtering in wavelet interframe codecs, called the unconstrained motion compensated temporal filtering (UMCTF). This framework allows flexible and efficient temporal filtering by combining the best features of motion compensation, used in predictive coding, with the advantages of interframe scalable wavelet video coding schemes. UMCTF provides higher coding efficiency, improved visual quality and flexibility of temporal and spatial scalability, higher coding efficiency and lower decoding delay than conventional MCTF schemes. Furthermore, UMCTF can also be employed in alternative open-loop scalable coding frameworks using DCT for the texture coding.

1. INTRODUCTION

Wavelet video coding schemes can provide flexible spatial, temporal, SNR and complexity scalability with fine granularity over a large range of bit-rates, while maintaining a high coding efficiency. Early contributions to the field of wavelet and multi-resolution video coding were provided, among others, by Gharavi [1], Zhang and Zafar [2], Taubman and Zakhor [3]. Advances in wavelet image compression have also significantly influenced wavelet video coding. Said and Pearlman [4] introduced Set Partitioning in Hierarchical Trees for efficient image coding that was later extended by Kim *et al* [5] for 3D wavelet video coding.

In this paper we introduce a new framework for temporal filtering in wavelet interframe codecs called the unconstrained motion compensated temporal filtering (UMCTF) [6]. This framework allows flexible and very efficient temporal filtering by combining the best features of motion compensation used in predictive coding with MCTF. UMCTF involves designing temporal filters appropriately to enable greater flexibility in temporal scalability, while also improving coding efficiency by allowing greater adaptability to the video content.

This paper is organized as follows. We first describe conventional MCTF in Section 2 and highlight its problems and inefficiencies. We then introduce the general framework for UMCTF in Section 3. We describe different choices for the filters and the decomposition structures to enable various enhancements as compared to MCTF. Some UMCTF results are presented in Section 4 and the conclusions in Section 5.

2. CONVENTIONAL MCTF

MCTF was first proposed by Ohm [7] and later improved by Choi and Woods [8]. Unlike predictive coding, where decoded frames are used as references for the motion compensation of future frames, MCTF does not employ a temporal recursive structure. Instead, in MCTF encoding, the original frames are filtered temporally in the direction of motion and the temporally decorrelated signal is coded using 2D spatial wavelet transforms and embedded coding. In this conventional MCTF framework, successive pairs of frames are temporally filtered using a two-channel Haar filter-bank to create low-pass (L) and high-pass (H) frames, thereby removing the short-term dependencies between successive frames. The long-term temporal dependencies are removed by further decomposing the L-frames using a pyramidal or multi-resolution decomposition structure. In conventional MCTF, the same Haar filter-bank is used at all various temporal decomposition levels. MCTF-based wavelet coding provides many advantages over conventional motion compensation algorithms for predictive coding, such as providing flexible spatio-temporal-SNR and complexity scalabilities, as well as improved error resilience due to clear prioritization of the coded video coefficients. Nevertheless, despite its significant advantages, conventional MCTF suffers from the following constraints and inefficiencies.

- **Low efficiency temporal filtering.** Due to the presence of irregular motion of objects in the scene or scene changes, good matches cannot always be found using uni-directional motion-estimation as in MCTF. As a result of this, compensation and filtering are performed across poorly matched regions, leading to the creation of annoying visual artifacts in the L frames and reduced coding efficiency.
- **Low quality and constrained temporal scalability.** Since temporal scalability is achieved in MCTF-based interframe wavelet coding by transmitting only the L-frames associated with a specific frame rate, poor quality L-frames translate directly into low visual quality when the video is decoded at lower temporal rates. This also directly affects the visual quality for spatial scalability at lower temporal frame rates. Moreover, due to the temporal filtering in pairs, only dyadic (powers-of-two e.g. half, quarter, one-eighth) frame-rate scalability can be achieved.
- **Increased delay.** Conventional MCTF incurs a long delay whenever decoding the video at full frame rate, due to the low-pass filtering of the frames at the various resolutions.

These inefficiencies are a direct consequence of the rigid temporal filtering methods in MCTF, i.e. fixed filter choice, fixed number of levels etc. To solve these conventional MCTF inefficiencies, we propose using more flexible temporal filtering, such as adapting the decomposition structure, number of decomposition levels, filter choices etc.

3. UNCONSTRAINED MCTF

We introduce a new flexible framework for motion compensated temporal filtering called Unconstrained MCTF (UMCTF) that provides higher coding efficiency, improved visual quality and flexibility of temporal and spatial scalability, and lower decoding delay.

3.1. Notation

First, the notation used subsequently is introduced.

0-7803-7663-3/03/$17.00 ©2003 IEEE

This paper was originally published in the Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, & Signal Processing, April 6-10, 2003, Hong Kong (cancelled). Reprinted with permission.

N : Number of frames in GOF temporally filtered together;

D : Number of levels in temporal decomposition pyramid; (the frames at level $D = 0$ are the original frames)

N^d : Number of frames at level $d \in [0, D]$

A_i^d : Unfiltered frames at level $d \in [0, D]$, $i \le N^d - 1$

L_i^d : Low-pass filtered frames at level $d \in [0, D]$, $i \le N^d - 1$

H_i^d : High-pass filtered frames at level $d \in [0, D]$, $i \le N^d - 1$, $H_0 = B$

P_i^d : Generic Picture, i.e. A/L, at location *i* in level *d*

M^d : Number of successive H frames at level $d \in [0, D]$ + 1 = Gap between successive L frames

f_i^d : High-pass filter used to create H_i^d frames. $i \le N^d - 1$.

g_i^d : Low pass filter used to create L_i^d and A_i^d frames, $i \le N^d - 1$. For A_i^d frames, $g_i^d(j) = \delta(i - j)$

$\left(v_{y,k\to i}^d, v_{x,k\to i}^d\right)$: Motion vector connecting frames *k* and *i* at level $d \in [0, D-1]$

$\hat{H}_i^d$: High-pass temporal filtered frames created by filtering H_i^d frames

$\hat{f}_i^d$: High-pass filter used to create $\hat{H}_i^d$ from H_i^d, $i \le N^d - 1$.

The above notation for UMCTF is illustrated in Figure 1.

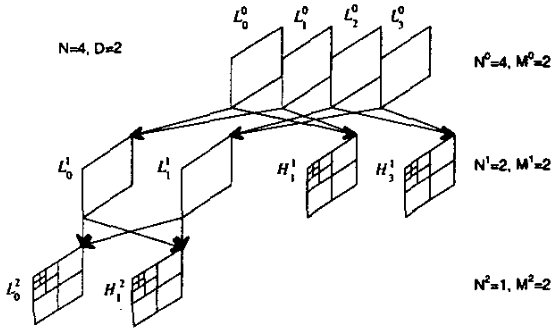

Fig. 1. Illustration of used UMCTF notation.

3.2. UMCTF framework

UMTCF provides adaptive temporal filtering through

- variable number of temporal decomposition levels based on the video content or desired complexity level;
- adaptive choice of filters enabling different temporal filtering enhancements;
- adaptive choice of filters, within and between temporal and spatial decomposition levels;
- variable number of successive H frames within and between levels, for flexible (non-dyadic) temporal scalability and temporal filtering enhancemements;
- different temporal decomposition structures.

These filters can be adapted across the different frames and between temporal levels, as well as within a frame, on a block or region level. Through appropriate choice of filters and decomposition structures many different improvements to MCTF become possible. For instance, predictive coding options such as sub-pixel accuracies, bi-directional prediction, multiple reference frames etc., may easily be introduced into the MCTF framework. Simultaneously, variable decomposition structures, such as modifying the number of decomposition levels, the number of successive H frames, decomposing H frames etc., can also be introduced.

3.3. Low-pass filters choices

To provide this flexibility while achieving perfect reconstruction, a complicated design and implementation of filters is necessary. Alternatively, UMCTF may use a very simple set of filters, the delta low-pass filter, obtained by setting $g_i^d(j) = \delta(i - jM^{d-1})$, i.e. leaving low-pass frames unfiltered. Once this choice is made, we may design the high-pass filters f_i^d without any constraints, to create H frames with the desired improvements, while guaranteeing perfect reconstruction. For instance, by appropriately choosing f_i^d, we can perform sub-pixel accurate, bi-directional, multiple reference temporal filtering etc.

Note that by setting $g_i^d(j) = \delta(i - jM^{d-1})$, the effective motion estimation and compensation methods used in predictive coding can also be introduced in MCTF. Nevertheless, UMCTF with this filter choice differs significantly from predictive coding. Specifically, in UMCTF we retain the multiresolution decomposition structure in order to exploit both long term as well as short term temporal dependencies. Also, we use a non-recursive prediction structure and fully embedded coding, such that spatial and SNR scalabilities do not suffer from the drift problems occuring in predictive coding. Most importantly, the flexibility and features supported by UMCTF are unmatched in predictive coding or conventional MCTF. We can adaptively change the number of reference frames, the relative importance attached to each reference frame, the extent of bi-directional filtering etc. Note that in the remainder of the paper, for simplicity, we will mainly exemplify the various temporal filtering enhancements using delta filters. Nevertheless, other low-pass filters (Haar, etc.) can also be employed.

3.3.1 Multiple Reference Frames in UMCTF

It has been shown in standards like H.26L that the use of multiple reference frames significantly improves the quality of matches obtained during motion estimation. Within the UMCTF framework, we can also introduce the multiple reference frames concept to interframe wavelet coding. The multiple reference temporal filtering[1] in UMCTF may be written as follows:

$$H_k^{d+1}(y, x) = \sum_{\substack{i=k-R^d,\\ i\ge 0}}^{k} \left(f_k^d(i) P_i^d\left(y + v_{k\to i}^y, x + v_{k\to i}^x\right)\right),$$

where the motion vector $\left(v_{y,k\to i}^d, v_{x,k\to i}^d\right)$ links source frame *k* to reference frame *i*, and R^d is the maximum number of allowable reference frames at temporal decomposition level *d*. Note that both A and H frames are represented by the generic picture '*P*' for ease of notation, and this includes the current frame *k* as well as the R^d used reference frames. With the

[1] We use the term 'filter' loosely. In this case the result is obtained by correlating our filter with the input frames and considering the result with no shift.

exception of $f_k^d(k)>0$, all other filter coefficients can be chosen appropriately to control the influence of a reference frame on the filtered result. For instance, if only the best reference frame is used during the filtering, then only one of the reference frames has a non-zero filter coefficient associated with it. The tradeoff between the improved prediction and the bits required for sending additional motion vectors may be exploited depending on the sequence characteristics, optimal bit rate versus quality etc.

3.3.2 Bi-directional Motion Estimation and Filtering

Bi-directional filtering can be used in conjunction with multiple reference frames filtering to further improve the motion estimation and temporal filtering process.

$$H_k^{d+1}(y,x)=\sum_{\substack{i=k-R_p^d,\\ i\geq 0}}^{k}\left(f_k^d(i)P_i^d\left(y+v^y_{k\to i},x+v^x_{k\to i}\right)\right)+\sum_{\substack{i=\left\lceil \frac{k}{M^d}\right\rceil\\ i<\frac{N}{M^d}}}^{\left\lceil \frac{k}{M^d}\right\rceil+R_f^d-1}\left(f_k^d\left(iM^d\right)P_i^d\left(y+v^y_{k\to iM^d},x+v^x_{k\to iM^d}\right)\right)$$

R_p^d is the maximum number of reference frames from the past, and R_f^d is the maximum number of reference frames from the future. While all frames from the past can be used as reference frames (including frames that are filtered into H frames at the current level), we limit the choice of reference frames from the future. Only frames that are filtered into L or A frames, at the current level, are used as reference frames from the future. This is done in order to avoid increasing the complexity and the delay. All the L and A frames are decoded before the H frames, so they can be used as references from the future without increasing the decoding delay. Moreover, to keep the delay limited, R_f should be kept small.

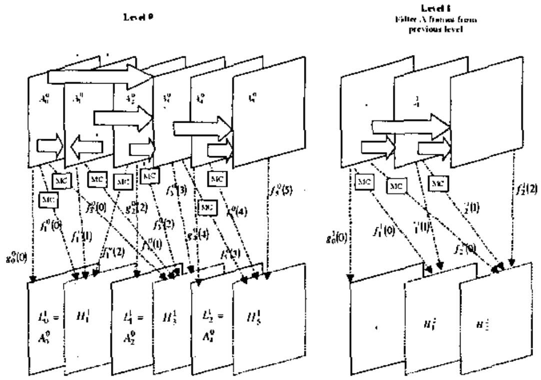

Fig. 2. Pyramidal Temporal Decomposition Scheme.

3.4. Variable Decomposition Structures

We now focus on the increased flexibility in terms of temporal and spatial scalability provided within the UMCTF framework. While with conventional MCTF, only dyadic frame-rate scalability can be achieved, with UMCTF unconstrained temporal scalability (i.e. any fraction of the full frame-rate) can be simply obtained as in the predictive coding case, by varying the number of H-frames between successive A/L-frames at the different temporal decomposition levels, i.e. the M^d. For instance, to achieve a sixth of the full-frame rate, we can use $N=6$, $D=2$ set $M^0=2$, and $M^1=3$. We show an this in Figure 2. For this example, we also set $R_p^0=3$, $R_f^0=1$, $R_p^1=2$ and choose $g_i^d(j)=\delta\left(i-jM^{d-1}\right)$.

3.5. H frames decomposition

In conventional MCTF schemes, the H frames are not temporally filtered and decomposed, based on the assumption that they do not retain any temporal redundancies. However, dependent on the sequence characteristics, this assumption is not always true. Hence, we can create $\hat{H}_i^d$ by performing motion estimation and filtering across the H frames using the high-pass filters $\hat{f}_i^d$. As there is a smaller amount of correlation between H frames as compared to that between A or L frame, the coding efficiency gains obtained by such decompositions are not as large as for the A or L frames and, the temporal pyramid can be terminated in this case after only one decomposition level. Furthermore, to limit the complexity and associated number of motion vectors, bi-directional filtering and no multiple reference frames are desirable in most practical implementations.

4. RESULTS

4.1. Results on Coding Efficiency

In this section, we evaluate the coding efficiency gain provided by the UMCTF framework, with $g_i^d(j)=\delta\left(i-jM^{d-1}\right)$. In the experiments, the sequences are CIF resolution, 30Hz. For the spatial transform and entropy coding, we used the EZBC method developed in [8]. For all the experiments in this section we used full search block motion estimation with 16×16 blocks, with half-pixel accuracy and a search range of $[-64,64]$. Moreover, we use the following basic settings for UMCTF: $N=16$, $D=4$, $M^d=2$ for all levels *d*. Figure 3 illustrates some sample results for the Foreman sequence using the following three coding options.

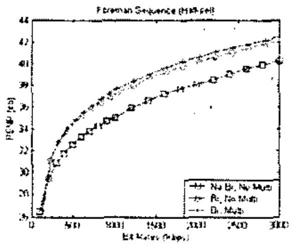

Fig. 3. PSNR results evaluating different coding options.

- **(No Bi , No Multi)**: No Bi-directional filtering and Multiple reference frames, $R_p^d=1$, $R_f^d=0$. The filter f_k^d has coefficients $f_k^d(k)=1$ and $f_k^d(k-1)=-1$ in this case, with all other filter coefficients being 0.
- **(Bi, No Multi)**: Bi-directional filtering, but No Multiple reference frames, $R_p^d=1$, $R_f^d=1$. Hence, $f_k^d(k)=1$ and $f_k^d(k-1), f_k^d(k+1)\in\{-0.5,0,-1\}$, where $f_k^d(k-1)+f_k^d(k+1)=-1$, with other filter coefficients being 0.
- **(Bi, Multi)**: Bidirectional filtering and Multiple reference frames, $R_p^d=N$, $R_f^d=1$, where only the best reference

frame was used during filtering. Hence, $f_k^d(k)=1$ and $f_k^d(j), f_k^d(k+1) \in \{-0.5, 0, -1\}$, with $f_k^d(j)+f_k^d(k+1)=-1$, where frame j provides the best match from the past for the current block of frame k. All other coefficients are set to 0.

The results improve significantly with the introduction of bi-directional filtering, despite the additional set of motion vectors that needs to be coded. Multiple reference frames further improve the results of the scheme with bi-directional filtering. For all of the following results reported, we use UMCTF parameter settings as for the (Bi, Multi) case. Subsequently, we compare UMCTF and the conventional MCTF scheme (MC-EZBC) in [8] using the same block based full search motion estimation, the same search range and the EZBC scheme for the spatial-domain texture coding.

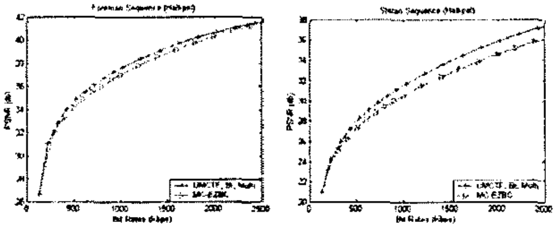

Fig. 4. PSNR results for UMCTF compared to MC-EZBC.

The results in Figure 4 show that UMCTF outperforms MC-EZBC, especially when the video has higher spatial detail and larger temporal motion. This dependence of the results on content is important, since it illustrates that higher efficiency can be obtained by making the filtering content dependent. In these experiments, a heuristic and sub-optimal rate allocation strategy was used for UMCTF. Due to our our use of non-orthonormal filters, we need to design a rate allocation strategy that addresses the relative importance of frames. As an example, if a frame is used as reference by many other frames, it is more important, and must be allocated more bits. Our pseudo-rate-allocation strategy weights each filtered frame using a weight equal to the number of times it is used as a reference by the other frames in the GOF. Hence, frames $\{A_0^4, H_1^4, H_1^3, H_3^3, H_1^2, H_3^2, H_5^2, H_7^2, H_1^1, H_3^1, H_5^1, H_7^1, H_9^1, H_{11}^1, H_{13}^1, H_{15}^1\}$ are weighted using the weights {16, 11, 13, 6, 12, 9, 6, 3, 8, 7, 6, 5, 4, 3, 2, 1}. A more accurate estimate of the relative importance may be obtained by summing all the filter taps applied to the current frame while filtering other frames, and this may be done at the block level or the frame level. Importantly, this weighting scheme is heuristic, and better results can be obtained with an improved rate allocation algorithm that is dependent on the video content etc. Hovewer, even with such a sub-optimal rate allocation strategy, UMCTF outperforms conventional MCTF.

4.2. Results on Temporal Scalability

To demonstrate the efficiency of UMCTF to support non-dyadic decompositions, and hence enable decoding of video at arbitrary fractions of the full frame rate, we compare the performance of the UMCTF in two different settings.

- $N=16$, $D=4$, $M^d=2$ for all levels d $R_p^d=N$, $R_f^d=1$.

We only use the best reference frame during filtering, i.e. $f_k^d(k)=1$ and $f_k^d(j), f_k^d(k+1) \in \{-0.5, 0, -1\}$, with $f_k^d(j)+f_k^d(k+1)=-1$, where frame j provides the best match from the past for the current block of source frame k. All other coefficients are set to zero. We label this as the **AHA scheme**.

- $N=9$, $D=2$, $M^d=3$ for all levels d $R_p^d=N$, $R_f^d=1$.

As before, we use only the best reference frames from the past during filtering. We label this as the **AHHA scheme**.

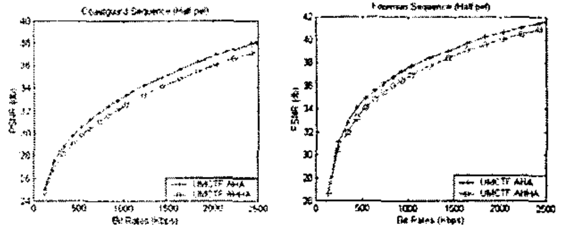

Fig. 5. Comparison of different decomposition structures.

As can be concluded from the results portrayed in Figure 5, the AHHA scheme has a loss in performance of 0.2~0.5 dB, since even though the H frames in the two cases are identical, the A frames get farther apart. Hence, the filtering of A frames at the higher temporal levels leads to worse prediction due to this increased distance and, as a result, poor filtering.

5. CONCLUSION

UMCTF provides a general and flexible framework that allows easy introduction of enhancements and features in temporal filtering for interframe wavelet coders. UMCTF may be used to integrate the best features of predictive coding techniques while retaining the significant advantages of MCTF. For instance, we can have multiple reference frames, arbitrary sub-pixel accuracy, and bi-directional filtering along with the lack of a prediction loop, no drift problems and truly scalable bitstreams. Importantly, by appropriately choosing the UMCTF "controlling parameters", easy adaptation to the desired video/network/device characteristics can be performed. Also, unconstrained temporal scalability can be easily provided, unlike in conventional MCTF. Furthermore, by choosing delta filters as the low-pass filters for UMCTF, the incurred delay can be considerably minimized as compared to the conventional MCTF.

REFERENCES

[1] H. Gharavi, "Subband Coding of Video Signals" in *Subband Image Coding, Kluwer Academic Publishers*, 1991.

[2] Y.-Q. Zhang and S. Zafar, "Motion-compensated wavelet transform coding for color video compression", *IEEE Trans. CSVT*, vol. 2, no.3, pp. 285-96, Sept. 1992.

[3] D. Taubman and A. Zakhor, "Multirate 3-D subband coding of video", *IEEE Trans. Image Proc.*, vol. 3, pp. 572-588, Sept. 1994.

[4] A. Said and W. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees", *IEEE Trans. CSVT* vol. 6, pp. 243-250, June 1996.

[5] B.-J Kim, Z. Xiong, and W. A. Pearlman, "Low bit-rate scalable video coding with 3-D Set partitioning in Hierarchical Trees", *IEEE Trans. CSVT*, vol. 10, pp. 1374-1387, Dec. 2000.

[6] D. S. Turaga and M. van der Schaar, "Unconstrained motion compensated temporal filtering", MPEG Contrib.M8388, May 2002.

[7] J. R. Ohm, "Three-dimensional subband coding with motion compensation", *IEEE Trans. Image Proc.*, vol. 3, no. 5, Sept. 1994.

[8] S.-J. Choi and J. W. Woods, "Motion compensated 3-D subband coding of video", *IEEE Trans. Image Proc.*, vol. 8, no. 2, Feb. 1999.