Three Dimensional Wavelet Based Approach for a Scalable Video Conference System

Patrizio Campisi, Mauro Gentile, and Alessandro Neri

Dipartimento di Ingegneria Elettronica, Università degli Studi Roma Tre,
Via Della Vasca Navale 84, I 00146 Roma, Italy
E-mail: (campisi,neri)@ele.uniroma3.it

Abstract

In this paper a video coder based on a spatio-temporal multiresolution pyramid generated by a 3D separable Wavelet transform (WT) is presented. The aim of the proposed coding scheme is to provide an effective scalability of the user profile by organizing the full resolution encoded bit stream into substreams, one for each temporal layer of the pyramid, which are then partitioned into substreams at increasing spatial resolution. This allows to simultaneously distribute, at different time and space resolutions, the same video to multiple users accessing the videoconference services through channels of different capacity, without multiple encoding.

1. Introduction

Dynamic access to telecommunication services, involved by the personal communication paradigm, implies a more efficient use of channel capacity, especially when broad-band services, like videoconference, are requested by mobile users. In fact, in next generation videoconference systems it is foreseen that the various attendants will be connected to a central multipoint video distribution center through both fixed and mobile channels of different nature and capacity.

However current video coder-decoder (codec) architectures based on block oriented DCT present intrinsic limitations with respect to dynamic spatial and temporal scalability of the quality of service. Thus, in order to avoid the use of a parallel bank of coders, current solutions are based on the distribution of a unique bit stream, carrying the encoded video, whose nominal bit-rate is set according to the capacity of the worst connection. As a consequence, channel resources are often under employed, and an insufficient quality, with respect to expectation, is perceived by those users accessing the videoconference network through fast channels.

In this contribution we propose a video coder architecture, based on a 3-D separable Wavelet transform (WT), consisting of a spatial 2D WT cascaded with a 1D temporal Haar WT, specifically designed to meet both spatial and temporal scalability requirements.

Space-time wavelet based coders have been already introduced in [1], [2], [3], [4].

The aim of the proposed coding scheme, is to provide an effective scalability of the user profile by organizing the full resolution bit stream into substreams, one for each temporal resolution which are then partitioned into substreams at increasing spatial resolution, and transmitting to each user a subset of substreams compatible with the capacity of the connecting channel.

The proposed method is based on the extension to the temporal axis of the zero-tree coding algorithm of the wavelet coefficients, originally proposed by Shapiro for still images in [5] and successively improved by a different implementation based on set partitioning in hierarchical trees (SPIHT) [6].

In fact the statistical analysis performed on several sequences evidenced that, although the wavelet transform approximates quite well an ideal "decorrelator", a statistical dependency between wavelet transform coefficients still remains. Therefore the 3D WT of video sequences presents the same persistence properties across different scales, both spatial and temporal, already observed in still images [5]. This in turn implies that, in the spatio-temporal WT domain, the wavelet coefficients across scales tend to coagulate in correspondence of visually significant details such as motion of edges, corners, lines etc. so that to null WT coefficients at lower spatio-temporal resolution correspond null WT coefficients at higher spatio-temporal resolution.

Thus the WT coefficients can be efficiently encoded following in an ordered manner the tree ramifications (sub-trees) starting from the lowest spatio-temporal

0-7803-5467-2/99 © 1999 IEEE
resolution plane of the wavelet coefficients.

The paper is organized as follows: in section 2 the bit stream organization is presented. In section 3 motion compensation in the wavelet domain is addressed. The spatio-temporal extension of the zero-tree coding technique is then outlined in section 4. Finally the experimental results and the conclusions are discussed in section 5.

2. The bit stream organization

A video conference system broadcasts the signal to users who are connected to a Multipoint Conference Unit (MCU) through channels of different capacity. As mentioned in the introduction, in DCT based coders, the frame rate, the image width, and the DCT quantization are set in order to allow the transmission through the worst connection. This sub-optimal choice is suggested by the attempt of avoiding expensive multiple coding-decoding devices, one for each capacity channel. The drawback of this architecture is that users connected to the MCU through fast channels are forced to receive the data at a bit rate lower than the maximum they can support thus leading to a poorest video quality.

To circumvent this limitation, the MCU should be designed to generate a unique data-stream composed by a “basic” sub-data-stream and by additional sub data-streams which, when added to the basic one, would allow to reconstruct, at the decoder side, a video with increased either both spatial or/and temporal resolution.

Therefore the conference attendees could get, from the MCU, the needed sub data-streams which, once decoded, would allow them to reproduce the video at the highest quality compatible with their access channel capacity. This approach leads to a “scalable” coding procedure.

In order to illustrate how the multiresolution decomposition can be utilized to fulfill this task, let us first examine the spatial scalability requirement. In this case, given a video sequence \( I = \{ I[m], \ m = 1, \ldots, M \} \) we can apply to each frame a 2D, separable, multiresolution analysis [7].

More specifically, given the approximation \( I^{(r+1)}[m] \) at resolution \( r+1 \) of a frame \( I[m] \), let us denote with \( \{ \mathcal{J}_i^{(r)}, \ i = \text{hor, vert} \} \) the projection operators acting respectively along the horizontal and vertical directions, so that \( \mathcal{J}_i^{(r)} I^{(r+1)}[m] \) is the coarser approximation at scale \( r \) along the corresponding direction. In addition let \( \{ \mathcal{K}_i^{(r)}, \ i = \text{hor, vert} \} \) be the projection operator orthogonal to \( \mathcal{J}_i^{(r)} \), so that

\[
I^{(r+1)}[m] = \mathcal{J}_i^{(r)} I^{(r)}[m] + \mathcal{K}_i^{(r)} I^{(r)}[m]
\]

Consequently, the spatial separable 2D decomposition of the frame \( I[m] \) is defined through the recursion:

\[
\begin{align*}
I^{(N)}_{LL}[m] &= \mathcal{J}_{\text{hor}}^{(r)} \mathcal{J}_{\text{vert}}^{(r)} I^{(r+1)}[m] \\
I^{(r)}_{LH}[m] &= \mathcal{J}_{\text{vert}}^{(r)} \mathcal{K}_{\text{hor}}^{(r)} I^{(r+1)}[m] \\
I^{(r)}_{HL}[m] &= \mathcal{K}_{\text{vert}}^{(r)} \mathcal{J}_{\text{hor}}^{(r)} I^{(r+1)}[m] \\
I^{(r)}_{HH}[m] &= \mathcal{K}_{\text{vert}}^{(r)} \mathcal{K}_{\text{hor}}^{(r)} I^{(r+1)}[m]
\end{align*}
\]

with initial condition

\[
I_{LL}^{(R_0)}[m] = I[m]
\]

where \( R_0 \) denotes the finest resolution.

Consequently, each subset \( (2)-(5) \) of the decomposition coefficients can be separately encoded, so that a pyramidal organization of the encoded bit-stream can be easily introduced, for each frame of the sequence.

While this approach is effective for separate transmission of still images, removal of the temporal redundancy present in video sequences requires a joint processing of a group of frames (GOF), whose number is practically limited by the maximum allowed delay. In this case, an additional multiresolution decomposition along the temporal axis can efficiently compress the original data, and fulfill, at the same time, the scalability requirements.

For sake of simplicity, let us begin the illustration of how such a technique can be devised by analyzing the simple case of quasi-static video, in the sense that interfame modifications induced by object motion are practically negligible.

At this aim, let us denote with \( I_{i,j}^{(r)}[m](p,q) \) the multiresolution decomposition coefficient of the \( (i,j) \) component \( (i=H.L \text{ and } j=H.L) \) corresponding to the space coordinates \( (p,q) \) at scale \( r \) of the \( m \)-th frame. Given a group of \( n_{GOF} \) frames, these coefficients are extracted and collected into the array \( \{ I_{i,j}^{(r)}[m](p,q), \ m = 1, \ldots, n_{GOF} \} \) to which a 1-D temporal multiresolution analysis is applied.

As outlined in Fig.1, where the case of a space-time multiresolution analysis using \( L_s=2 \) spatial levels and \( L_t=1 \) time levels for the space-time pyramid is outlined, the data stream can be properly organized in a pyramidal way in layers, one for each temporal resolution plane and each of them into sub-layers, one for each spatial level, so that the user can select from the stream the information related to desired space and time resolution.

Specifically, the data contained in the substream labeled with “A” allow to reproduce the GOF at a spatial resolution reduced by two and at a temporal resolution reduced by one. The data in substream “B” allow to reproduce the GOF at half the nominal resolution in both space and time, the data in the substream “C” allow to
reproduce the GOF at the full spatial resolution but at the temporal resolution reduced by one. Finally the whole stream "D" allow to reproduce the GOF at full resolution.

Let us note that, in practice, compact support temporal projection operators, let say \( f_{\text{temp}}(s) \), \( f_{\text{time}}(s) \), as the Haar wavelets, have to be employed in order to cope with the maximum delay constraint.

3. Motion compensation in the wavelet domain

As further discussed in the next section, one of the main characteristics of the wavelet transform is its ability to compact, or focus, the signal energy into the lower resolution layers of the pyramid. This property, that, together with the coefficient sparseness, is the corner stone of the effectiveness of wavelet based image coders, is weakened by interframe variations induced by object motion.

To reduced the impact in coding efficiency caused by object motion as well as by changes in the video imaging system (e.g. pan, scroll, zoom) we propose to interleave each stage of the temporal Haar WT decomposition with a motion estimation and a realignment performed on the set of WT coefficients.

This procedure is somehow similar to the motion compensation employed by DCT based, block oriented, coders, that to efficiently remove the temporal redundancy apply the DCT transform to the difference between the current block and that block of the previous reference frame giving the best match.

More in detail, in our scheme, once the spatial wavelet transform (SWT) on the frames belonging to the GOF under analysis is computed, wavelet coefficients corresponding to the same spatial scale are partitioned into small blocks and the blocks are realigned using a block matching procedure.

As illustrated in Fig.2 where the block diagram of the proposed coder is depicted, to preserve the temporal scalability a motion estimation and realignment precedes each stage of the temporal decomposition.

More in detail the WT coefficient realignment operates as follows. For a given pair of consecutive spatial WT's (SWT's) each resolution plane is partitioned into small blocks and two lists are created. The first one, named in the following as free block list, collects the blocks of the first SWT, acting as reference, and the second, named unassociated block list, contains the blocks of the other SWT whose elements have to be realigned. Then, for each block of the unassociated block list the best match in the free block list (i.e. the block corresponding to the lowest \( L^1 \) norm of their difference) is evaluated.

The norm of the residual error (i.e. the \( L^1 \) norm of the said difference) is then employed to sort the unassociated block list in ascending order. The first element of the reordered list is then moved into the list of realigned blocks together with the information about its best match, which is contemporary removed from the free block list. This procedure is iteratively applied to each element of the ordered list until either the best match of the current block has been already utilized as reference by a previous block (obviously with a smallest residual error) or the list is empty. When the matching block has already been utilized and therefore removed from the free list, the whole algorithm is restarted and reapplied to the current lists of unassociated and free blocks.

Let us note that this competitive association rule guarantees that all blocks of the reference SWT are utilized. On the other hand, one of the drawbacks of the block competition is that during the final iterations of the algorithms the residual errors can become too high. This effect can be mitigated by introducing horizontal and vertical split of unassociated blocks. In other words, when residual error associated to the best match exceeds a predefined threshold, horizontal and vertical splits are iteratively applied in the attempt of reducing it.

A reduction of the computational load can be
obtained by assuming as negligible (i.e. practically null) the motion of small energy blocks.

On the other hand, the high correlation across scales presented by the WT coefficients, reflects into a high correlation between WT coefficient displacements needed to compensate for actual motion at different scales, so that the information sent to the decoder in order to properly realign the WT coefficients can be organized into progressive refinements of the displacement utilized at the lower spatial resolution.

In other words, we can code only the increments of the motion vector in the sub-band under analysis with respect to the one estimated for the same block at the lower spatial resolution. This leads to a bit saving in representing the motion estimation between blocks of homologous sub-bands.

4. Entropy coding of spatio-temporal zero-trees

Entropy coding of the multiresolution bit stream can be achieved by extending the space-time three dimensional case the embedded zero-tree coding approach developed by Shapiro in [5] for still images which uses a discrete pyramidal wavelet decomposition. The rationale behind Shapiro’s algorithm relies on the property, observed on a large amount of images of interest, that when a wavelet coefficient (called parent), at a coarse scale, is negligible, then its children, i.e. the wavelet coefficients at the finer scales related to the same spatial locations, are likely to be negligible as well. As proposed by Shapiro, this behavior can be fruitfully utilized by coding the wavelet coefficients by means of an ordered exploration of the space-scale quad-tree.

In fact, due to the sparseness of the WT coefficients, at each node a descriptor is used to inform the decoder whether the sub-trees originating from the current node present relevant nodes or if its exploration can be stopped due to their absence. Specifically three situations are considered:

- zero-tree (ZT): the WT coefficient associated to the current node as well as the WT coefficients of the sub-tree’s nodes whose root is the current node are insignificant and therefore can be skipped;
- significant (S): the magnitude of the WT coefficient associated to the current node exceeds a predefined threshold. The corresponding value is quantized and entropically encoded, then the ordered exploration of the sub-tree originating from the current node is performed;
- isolated zero (IZ): the WT coefficient associated to current node is insignificant but significant nodes exist in the sub-tree originating from the current node, which is therefore explored.

However, as illustrated by Fig. 3, where the magnitudes of the WT coefficients (pseudo-images of the space-scale pyramids) at four temporal resolutions corresponding to the operators $RR$, $RH$, $HR$, and $HH$, of a GOF of four frames extracted from the “Mother and daughter” sequence are displayed, 3D WT’s of video sequences exhibit similar properties of clustering within the same scale and persistence across different scales of the WT coefficients that can be exploited in order to efficiently encode the bit-stream.

Specifically, let $c$ be a WT coefficient, and let $d_1$, $d_2$, $d_3$, $d_4$ respectively be the (eventually empty) subsets composed of nodes of the four spatial sub-trees having $c$ as root and respectively belonging to the ranges of $RR$, $RH$, $HR$, and $HH$. Then, for each non-empty subset we can use a Shapiro’s descriptor to instruct the decoder about when and how to continue to visit the sub-tree. Obviously, the descriptor sequences are entropically encoded by means of either the Huffman or the Lempel-Ziv algorithm. In addition, to account for the differences in statistical distribution of the WT coefficients vs. both the temporal and the spatial resolution, different subsets of descriptors are employed for the roots of each temporal and spatial sub-band.

Thus, the bit-stream is now composed of a main list consisting in the ordered sequence of the encoded sub-tree visiting rule descriptors carrying information on the location of significant WT coefficients, and an associated list containing the quantized coefficients. The quantization step can be iteratively refined to meet the required bit rate.

Since the block displacement utilized for the realignment of SWT blocks, as specified in the previous section, exhibit a similar sparseness the zero tree method
5. Experimental results and concluding remarks

The effectiveness of the method in achieving scalability of the user profile is illustrated in Fig. 4 where a few frames of “Mother and daughter” at different resolutions extracted from the same encoded bit-stream are shown. Since, usually, an overall constraint on the maximum bit-rate is specified, temporal and spatial resolutions can be jointly traded off based on the detection of dynamic events or on the density of spatial details, by acting on the WT quantization steps. In fact interleaving the temporal WT with the motion compensation allows to decode a video sequence at any predefined space-time resolution without knowledge of the full resolution motion estimate as required by DCT based coders.

Let us finally observe that organization of the full resolution bit stream into substreams, one for each resolution layer of the space-time pyramid, can be directly applied even to video sequences of multimedia archives. Moreover in this case the maximum number of frames constituting the GOP is not limited by real-time constraints so that highly efficient browsing techniques can be devised.

References