MODELING MOTION FOR SPATIAL SCALABILITY

Nikola Božinović and Janusz Konrad

Department of Electrical and Computer Engineering, Boston University
8 St. Mary St., Boston, MA 02215

ABSTRACT

The dramatic proliferation of visual displays, from cell phones, through video iPods, PDAs, and notebooks, to high-quality HDTV screens, has raised the demand for a video compression scheme capable of decoding a "once-encoded" video at a range of supported video resolutions and with high quality. A promising solution to this problem has been recently proposed in the form of wavelet video coding based on motion-compensated temporal filtering (MCTF); scalability is naturally supported while efficiency is comparable to state-of-the-art hybrid coders. However, although rate (quality) and temporal scalability are natural in mainstream "t+2D" wavelet video coders, spatial scalability suffers from drift problems. In the light of the recently proposed "2D+t+2D" modification, which targets spatial scalability performance, we present a framework for the modeling of spatially-scalable motion that is well matched to this new structure. We propose a motion estimation scheme in which motion fields at different spatial scales are jointly estimated and coded. In addition, at lower spatial resolutions, we extend the block-wise constant motion model to a higher-order model based on cubic splines, effectively creating a "mixture motion model" that combines different models at different supported spatial scales. This advanced spatial modeling of motion significantly improves the coding efficiency of motion at low resolutions and leads to an excellent overall compression performance; spatial scalability performance of the proposed scheme approaches that of a non-scalable coder.

1. INTRODUCTION

In view of excellent performance of the wavelet-based JPEG-2000 still-image compression standard, the growing demand for efficient scalable video coding has brought, over the last decade, the emergence of a new coding paradigm based on three-dimensional discrete wavelet transforms (3D DWT). The structure of a typical 3D DWT coder is illustrated in Fig. 1, where T, S, and E denote temporal, spatial, and entropy blocks, respectively. This is often referred to as a "t+2D" scheme; the input video frames are first temporally analyzed, then spatially decomposed, and finally an entropy coder is used to encode spatio-temporal wavelet subbands and produce the output bitstream. This processing order is reversed in the decoder.

In the earliest attempts, a separable 3D DWT was applied to an image sequence with no motion compensation before the transform coefficients were entropy coded [1]. Global motion compensation was subsequently introduced in temporal filtering [2], followed by local, block-based motion compensation [3]. Motion compensation created problems for temporal DWT implemented in classical (transversal) way, destroying it's perfect reconstruction property for sub-pixel motion. As a solution, lifting implementation of temporal DWT was introduced [4, 5, 6] that guarantees perfect reconstruction of MC-DWT for arbitrary motion.

This work was supported by the National Science Foundation under grant CCR-0209055.

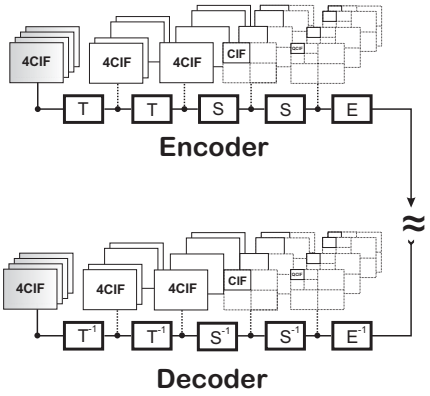

Fig. 1. Block diagram of a "t+2D" coding scheme

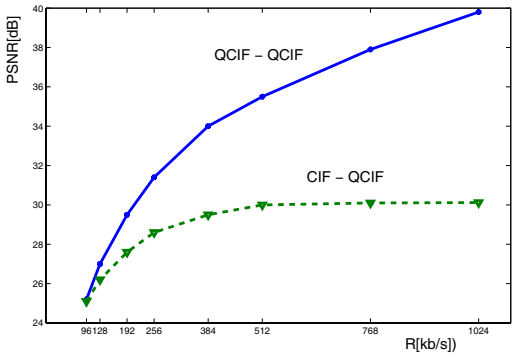

Fig. 2. R-D performance gap of a typical "t+2D" scheme for *Mobile* when decoding from different bitstreams: solid line – QCIF-encoded/QCIF-decoded, dashed line – CIF-encodied/QCIF-decoded (note the saturation due to the subband leakage problem).

2. SPATIAL SCALABILITY

Over the last few years, the issue of spatial scalability has gained particular prominence because of the variety of emerging displays supporting a wide range of resolutions. This is further compounded by the growth of high-definition TV; most of the future video material will be captured using high-resolution video cameras and its presentation on low-resolution devices will require spatial scaling. There are several solutions to this problem. First, video data can be transmitted at the highest spatial resolution and downconverted at the receiver, but this requires a high-bandwidth channel down to the very receiver. Alternatively, the full-resolution bit-stream can be transcoded to lower resolution (lower rate) at the interface of

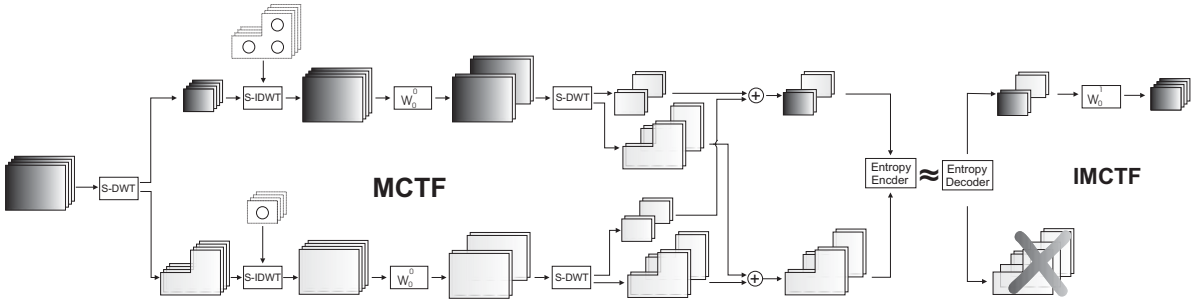

Fig. 3. Block diagram illustrating the subband leakage problem in a typical "t+2D" scheme.

different-bandwidth networks but this requires complex and costly transcoding-capable network switching gear. Another solution, used notably in video databases, is to encode and store the video material at different spatial resolutions (and, thus, different bit-rates) but this requires more complex data management as well as additional storage capacity. A better solution is to employ spatially-scalable video coding. The idea is to assemble a single bit-stream from which a sub-stream can be extracted and transmitted at lower bit-rate, and subsequently decoded at lower spatial resolution.

It was early recognized that spatial scalability can be easily embedded into "t+2D" wavelet coding schemes (Fig. 1). Such schemes show excellent performance when video is decoded at full spatial resolution, when the inverse MCTF in the decoder matches the forward MCTF in the encoder. However, the coding performance deteriorates when a decoder extracts video at reduced spatial resolution by omitting one or more levels of spatial synthesis. Not only are the visual quality and PSNR of such lower-resolution video inferior compared to the case of decoding this very resolution from matching-resolution bit-stream, but they also saturate even at high bit-rates (Fig. 2). Two major factors contributing to such poor performance of the "t+2D" structure are: 1. inappropriate motion modeling that creates visually annoying artifacts, and 2. subband leakage [7, 8, 9] induced by shift-variance of critically-sampled DWT transforms. In order to analyze the subband-leakage problem of a "t+2D" scheme, we consider equivalent structure (Fig. 3), where the input video sequence is separated into two subsequences (by means of spatial DWT analysis, zero-padding of appropriate subbands, and subsequent DWT synthesis). Both subsequences are then subject to MCTF (using the same motion), and spatial DWT.

It can be shown that MCTF of an image sequence containing only low spatial frequencies will produce high spatial frequencies in the output, while the same MCTF applied to image sequence with only high frequencies will create some low spatial frequencies at the output. This effect is is often referred to as *subband leakage*. When a "t+2D" decoder does not have access to the high spatial subbands, exact reconstruction of the original low spatial subband is not possible, even in absence of quantization errors.

The well-known "2D+t" schemes provide one solution for breaking this subband dependency; spatial synthesis in the decoder is performed last and does not affect other steps (in case all sub-bands are not decoded). Unfortunately, this comes at the cost of lower coding efficiency due to inefficient motion compensation in the critically-sampled wavelet domain. Recent solutions to the subband leakage problem proposed in the literature include shift of motion compensation to the overcomplete DWT domain [10], construction of better spatial filters (limiting the amount of spatial aliasing) [11], and optimal subband rate allocation [12], which assigns certain part of total bit-budget to otherwise discarded spatially-high DWT coefficients.

To improve the spatial scalability performance, we proposed a new MCTF design with motion fields optimized for each resolution level [13]. Concurrently and independently from our work, similar solutions appeared in the form of "2D+t+2D" schemes [7] and "In-Scale" MCTF schemes [14]. In "2D+t+2D" schemes, several levels of the so-called "pre-S" spatial DWT decomposition precede the *subband-independent* temporal filtering stage (Fig. 4); additional levels of spatial 2D-DWT decomposition (used strictly to increase coding efficiency) complete subband analysis. We use subscript/superscript index of motion field $\mathcal{W}$ to denote spatial/temporal resolution at which MCTF is performed. Note that MCTFs at all but the lowest spatial resolution are not performed in the transform domain – each modified temporal processing block ("Mod.-MCTF" in Fig. 4) consists of zero padding, inverse 2D-DWT, MCTF, and forward 2D-DWT.

The novelty of this approach lies in the parallel implementation of MCTFs at different scales, which eliminates the encoder/decoder mismatch and solves the high-to-low subband leakage problem. The "2D+t+2D" design supports flexible multi-resolution motion estimation; this is an improvement over the classical "t+2D" scheme, where low-resolution motion is simply *derived* from the full-resolution motion field (by means of scaling and subsampling). In the next section, we present a new framework for spatially-scalable motion representation that is well matched to the hierarchical structure of spatially-scalable "2D+t+2D" coder.

3. MOTION ESTIMATION FOR SPATIALLY-SCALABLE WAVELET-BASED VIDEO CODING

Scalable (lossy) coding of motion became an issue only recently with the introduction of scalable wavelet video coders. Motion with a degree of coding error, not permitted in hybrid coders, is permitted in wavelet-based coders due to the open-loop nature of MCTF. So far, only rate scalability of motion has been considered in order to improve compression performance at lower bit-rates [15], by means of shifting a portion of the total bit-budget from motion to texture. However, simple bit-plane coding of motion parameters is highly sub-optimal and not suitable for motion compensation at lower spatial resolutions. Below, we present a general framework for the estimation of spatially-scalable motion fields.

Let $\boldsymbol{W}$ be a set of motion fields at all spatial scales that are supported by the spatially-scalable coder, $\boldsymbol{W} = \{\mathcal{W}_0, \mathcal{W}_1, ..., \mathcal{W}_{L-1}\}$, where L is the number of spatial scales. In the most general case, the formulation of joint motion estimation across all spatial scales is:

$$\min_{\boldsymbol{W}} J(\boldsymbol{W}), \qquad J = \sum_{l=0}^{L-1} c_l (E_l + \lambda_l R_l), \tag{1}$$

where c_l's are scale-normalization factors, and E_l, λ_l and R_l are the distortion (typically SAD), regularization factor, and motion rate at

resolution level l. Minimizing this cost function is a computationally challenging task. Instead, we propose to solve iteratively, starting from the highest spatial scale[1] $L-1$, a simpler cost function:

$$\min_{\mathcal{W}_l} J_l, \qquad J_l = E_l + \lambda_l R_l|_{\mathcal{W}_{l+1},\mathcal{W}_{l+2},...,\mathcal{W}_{L-1}}, \qquad (2)$$
$$l = L-1, ..., 1, \;\; \mathcal{W}_L \triangleq 0,$$

where $\mathcal{W}_k, k > l$ are motion fields already estimated at scales higher than the current scale l. Typically, R_l is the rate needed to *predict* motion at scale l from all previously-estimated motion fields. Practical design of this prediction model and its prediction contexts depends on the particular motion models used. We propose here two motion estimation schemes: one employing HVSBM (hierarchical variable-size block matching) at all spatial scales, and the other using mixture of hierarchical cubic splines (at lower scales) and HVSBM.

3.1. Spatially-scalable motion estimation using HVSBM

Motion vectors in HVSBM are typically predictively coded by sequentially following the quad-tree decomposition structure. The prediction residual is then coded using the context-based adaptive arithmetic coding (CABAC). Previous attempts to generate scalable layered motion were usually limited to a single spatial resolution; large λ was used to generate motion base layer and progressively smaller λ's were used for the refinement. In our method, both spatial and cross-resolution predictors are allowed (Fig. 5). This adaptive prediction scheme is deployed on a macroblock level. The decision on a particular prediction mode is controlled by the variance of motion estimates and current depth in partition tree. The most frequently used prediction modes are median, weighted median, and average of $\vec{A}$, $\vec{B}$, $\vec{C}$, $\vec{P}$ (Fig. 5), as well as simple prediction using $\vec{P}$. In addition, macroblock partitioning from the next lower resolution is used to initialize the state of four MBs at the current layer; while these initial sub-blocks may be subject to additional partitioning, their merging is not allowed. This results in an efficient motion coding. In order to speed up motion estimation at higher resolutions, (scaled) motion vectors estimated at lower resolution are used as search candidates.

3.2. Spatially-scalable ME using spline-based motion model

Most of the current MC-DWT coders still use block-based motion models inherited from the hybrid coding structure. The excellent performance of block matching in the hybrid context can be easily explained: block-based motion perfectly aligns with block-based decorrelating DCT transform used. In contrast to the hybrid scenario, there is a mismatch between the local support of the block motion model and a global (multi-scale) nature of spatial 2D-DWT. The deficiency of the local, zero-order model (block-constant) in describing the scene dynamics at high spatial scale motivated us to propose motion model based on hierarchical cubic splines [16]. With more degrees of freedom, this model is bound to perform a better motion compensation at high spatial scales than the block-constant one. We model the horizontal and vertical components of the motion field $\mathcal{W}$ by two-dimensional splines defined on a control grid Γ that is a sublattice of Λ ($\Gamma \subset \Lambda$), where Λ is a lattice [17] on which each sequence frame is defined. A motion (displacement) vector at $\boldsymbol{x}$ can be expressed as follows:

$$\mathcal{W}[\boldsymbol{x}] = \sum_{\boldsymbol{y}\in\Gamma} \vec{\gamma}[\boldsymbol{y}]\beta^{(n)}(\boldsymbol{x}-\boldsymbol{y}), \quad \boldsymbol{x} \in \Lambda, \qquad (3)$$

[1]The highest scale corresponds to the lowest resolution and vice versa

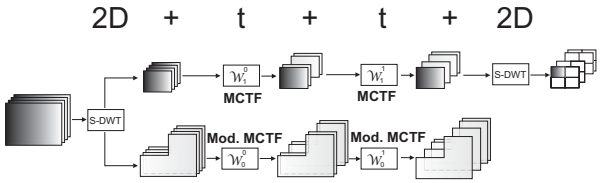

Fig. 4. "2D+t+2D" coding scheme supporting two levels of spatial scalability.

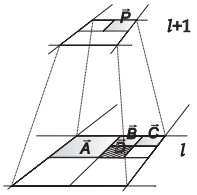

Fig. 5. Motion prediction: both, same-scale ($\vec{A}$, $\vec{B}$, $\vec{C}$) and previous-scale ($\vec{P}$) motion vectors are used to predict motion vector of a current block ($\vec{D}$).

where $\vec{\gamma}$'s are (vector) spline coefficients defined on Γ, while $\beta^{(n)}(\boldsymbol{x})$ is a 2-D separable spline basis function of order n. In order to compute spline coefficients $\vec{\gamma}$, we minimize (2) with respect to these coefficients using a variant of the Levenberg-Marquardt iterative nonlinear minimization [18].

We create a "mixture motion model" by replacing HVSBM with splines at lower layers of the spatially-scalable motion. A small number of spline control nodes can accurately describe motion field at high scale, resulting in improved motion compensation, lower motion rate, and better prediction of motion at the next spatial level. At the "switch" layer, the motion prediction scheme is similar to that of HVSBM (Fig. 5) – the only difference is that the predictor $\vec{P}$ from the previous layer $l+1$ is now derived from a spline-based field and may vary over the MB support. We note that, in contrast to HVSBM, the continuous spline model does not limit the precision of a cross-scale (low resolution) predictor; instead of a simple doubling of the motions vectors, better predictor is usually available when the continuous spline is re-evaluated at the current scale.

4. EXPERIMENTAL RESULTS

We implemented the proposed algorithms within the framework of the MSRA [19] scalable wavelet coder. For HVSBM, we used both spatial and cross-scale motion prediction, 16×16 initial macroblocks, and $1/8$-pel motion accuracy at each resolution level. For spline-based motion estimation we used only fixed-size control grid (8, 16, or 32 pixels) within a single spatial scale. Motion from different temporal resolutions was independently coded; we plan to work on a more efficient temporal motion coding in the future.

We first compare performance of the "2D+t+2D" coder using layered HVSBM motion (inputs are CIF resolution *Foreman* and *Mobile*) with that of a standard "t+2D" coder using non-scalable motion and same resolution encoding/decoding. It is clear that these two schemes are identical at the lowest supported resolution (QCIF). They both significantly outperform CIF-encoding/QCIF-decoding of the "t+2D" scheme (coding gain is the same as that in Fig. 2). The penalty introduced by layered motion coding and "2D+t+2D" structure is assessed at the full-resolution (Fig. 6): coding loss is less than 0.8dB for *Foreman* and 1dB for *Mobile*.

Next, we investigate the effects of spline-based motion layers. We compare full-resolution decoding performance of schemes with K levels of spline-based motion at the lowest spatial scales, and

$L - K$ levels of HVSBM at higher scales (no splines corresponds to $K = 0$, which is the standard all-block-matching approach; pure spline model is defined with $K = 3$). We use "2D+t+2D" coder configuration supporting three spatial resolutions: CIF, QCIF, and QQCIF. For *Foreman* and *Mobile*, the best results are obtained for $K = 1$ (Fig. 7), justifying advanced spatial motion modeling at higher spatial scales; performance of "2D+t+2D" scheme using layered "mixture motion model" approaches that of a single-resolution "t+2D" scheme using non-scalable motion.

5. CONCLUSIONS

In this paper, we motivated and proposed a framework for the estimation of spatially-scalable motion for fully-scalable wavelet video coding with the special emphasis on spatial scalability. We introduced a mixture motion model that combines HVSBM and cubic spline motion models at different supported spatial scales. A cross-resolution prediction and improved spatial modeling of motion lead to better spatial-scalability performance of "2D+t+2D" schemes without a significant coding loss at full resolution. We plan to develop a layered motion structure with more than one layer at a single spatial resolution, and work on the related rate-distortion optimization (layer-switching) problem.

6. REFERENCES

[1] G. Karlsson and M. Vetterli, "Three-dimensional subband coding of video," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, Apr. 1988, vol. 2, pp. 1100–1103.

[2] D. Taubman and A. Zakhor, "Multirate 3-D subband coding of video," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 572–588, Sept. 1994.

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

[4] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, 2001, pp. 1793–1796.

[5] L. Luo, J. Li, S. Li, and Y.-Q. Zhang, "Motion compensated lifting wavelet and its application in video coding," in *Proc ICME, Tokyo, Japan*, Aug. 2001.

[6] A. Secker and D. Taubman, "Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting," in *Proc. IEEE Int. Conf. Image Processing*, 2001, pp. 1029–1032.

[7] N. Mehrseresht and D. Taubman, "A flexible structure for fully scalable motion compensated 3D-DWT with emphasis on the impact of spatial scalability," *IEEE Trans. Image Process.*, (submitted).

[8] R. Xiong, J. Xu, F. Wu, S. Li, and Y.Q.Zhang, "Spatial scalability in 3D wavelet coding with spatial domain MCTF encoder," in *Proc. Picture Coding Symposium (PCS)*, Dec. 2004.

[9] N.Adami, M.Brescianini, M.Dalai, R.Leonardi, and A.Signoroni, "A fully scalable video coder with inter-scale wavelet prediction and morphological coding," in *Proc. SPIE VCIP*, July 2005.

[10] Y. Andreopoulos, M. Van Der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and J. Cornelis, "Complete-to-overcomplete discrete wavelet transforms for fully-scalable video coding with MCTF," in *Proc. SPIE VCIP*, July 2003, pp. 719–731.

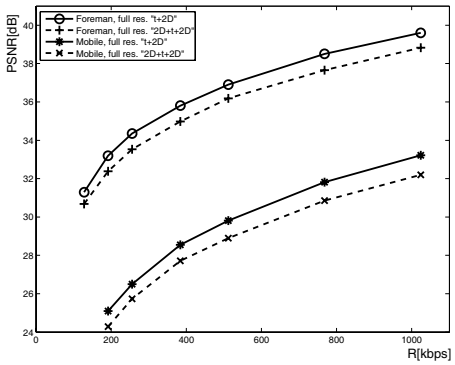

Fig. 6. "t+2D" vs. "2D+t+2D" at full resolution

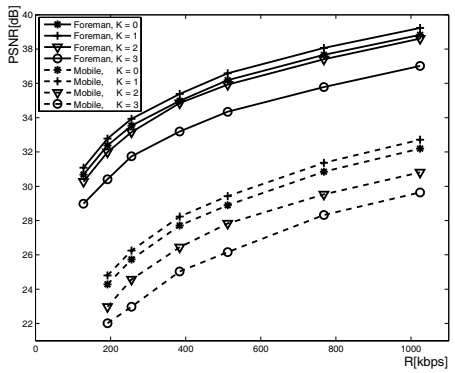

Fig. 7. R-D performance as a function of the number of spline-based motion layers (K). The best results are obtained for $K = 1$.

[11] Y. Wu, *Fully scalable subband/wavelet video coding system*, Ph.D. thesis, RPI, ECSE Dept., Sept. 2005.

[12] R. Xiong, J. Xu, F. Wu, and S. Li, "Optimal subband rate allocation for spatial scalability in 3D wavelet video coding with motion aligned temporal filtering," in *Proc. SPIE VCIP*, July 2005.

[13] N. Bozinovic, *Advanced motion modeling for 3D video coding*, Ph.D. Thesis Prospectus, BU ECE Dept., Nov. 2004.

[14] R. Xiong, J. Xu, F. Wu, and S. Li, "Studies on spatial scalable frameworks for motion aligned 3D wavelet video coding," in *Proc. SPIE VCIP*, July 2005.

[15] A. Secker, *Motion-adaptive transforms for highly scalable video compression*, Ph.D. thesis, University of New South Wales, School of Electr. Eng. and Telecom., Aug. 2004.

[16] R. Szeliski and H.-Y. Shum, "Motion estimation with quadtree splines," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 12, pp. 1199–1210, Dec. 1996.

[17] E. Dubois, "The sampling and reconstruction of time-varying imagery," Tech. Rep. 84–34, INRS Télé, Oct. 1984.

[18] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical recipes in C: The art of scientific computing*, Cambridge University Press, 2-nd edition, 1992.

[19] J. Xu, Z. Xiong, S. Li, and Y.-Q. Zhang, "Three-dimensional embedded subband coding with optimized truncation (3-D ESCOT)," *Appl. and Comp. Harmonic Analysis: Special Issue on Wavelet Applications*, vol. 10, May 2001.