A MODEL-BASED MOTION COMPENSATED VIDEO CODER WITH JPEG2000 COMPATIBILITY

Marco Cagnazzo [1,2]*, Thomas André* [1]*, Marc Antonini* [1]*, Michel Barlaud* [1]

[1]I3S Laboratory, UMR 6070 of CNRS, University of Nice-Sophia Antipolis
Bât. Algorithmes/Euclide, 2000 route des Lucioles - BP 121 - 06903 Sophia-Antipolis Cedex, France
Phone: 33(0)4.92.94.27.21 - Fax: 33(0)4.92.94.28.98 - {cagnazzo,andret,am,barlaud}@i3s.unice.fr

[2]Dipartimento di Ingegneria Elettronica e delle Telecomunicazioni,
Università Federico II di Napoli, via Claudio 21 - 80125 Napoli, Italy

ABSTRACT

We present a highly scalable wavelet-based video coder, featuring a scan-based motion-compensated temporal wavelet transform (WT) with lifting schemes which have been specifically designed for video. Output bitstream is compatible with JPEG2000, as it is used to compress temporal subbands (SBs). Rate allocation among SBs is done by means of an optimal algorithm which requires SBs rate-distortion (RD) curves. We propose a model-based approach allowing us to compute these curves with a considerable reduction in complexity. The use of temporal WT and JPEG2000 guarantees high scalability.

1. INTRODUCTION

Compression is a mandatory step in almost all applications of digital video, and so, since the late 80s, there has been a great interest towards video compression problems. Now several efficient video coding standards exist, all based on Motion Compensation (MC) and Discrete Cosine Transform (DCT) or its variations, which have been the enabling technologies for digital video, above all with the MPEG2 standard. These techniques keep to be improved: the new H.264 standard allows a considerable bit-rate saving for the same quality with respect to MPEG2.

Nevertheless, alternative techniques for video compression have been investigated for several years, as some problems are still far from being completely resolved. In particular, much attention has been reserved to Wavelet Transform (WT), which proved to be clearly superior to DCT for still image coding [1], and is employed in the recent JPEG2000 standard [2]. Moreover, WT offers a natural support to scalability, which is an imperative feature for video encoders aiming to be used over heterogeneous networks like the Internet. For these reasons, since the middle of 90s, WT-based video coders have been intensively studied [3, 4], but only recently they proved to be as effective as DCT-based video compression schemes. This is mostly due to Motion Compensated Lifting approach [5, 6, 7], used to reduce temporal redundancy of the input sequence.

We propose in this paper a WT-based video coder, which is fully compatible with JPEG2000. The coder performs firstly a temporal Motion Compensated WT, with a scan-based approach [8]. This temporal stage is described in section 2. Afterwards, temporal subbands (SBs) are encoded with JPEG2000. Coding resources are allocated to SBs with a model-based algorithm (described in section 3) which achieves optimal rate allocation by estimating the Rate-Distortion (RD) curves of each SB with a fast, spline-based method. In section 4, experimental results show that the proposed algorithm reduces remarkably the complexity of the RD curves estimation. Moreover, the respective performances of our encoder and of the new standard H.264 are equivalent.

2. TEMPORAL ANALYSIS

In this section, we describe the motion-compensated scan-based [8] temporal transform used in the first step of our coding algorithm.

2.1. Motion compensated temporal filtering

Motion compensation is the key to an efficient video coding scheme, as it leads to a substantial energy reduction in the high-frequency subbands by applying the transform along adapted motion trajectories. However, since motion compensation is a non-linear process, it cannot be implemented directly into a regular wavelet transform, unless the transform becomes non invertible. The use of lifting schemes leads to a fully invertible motion-compensated transform.

PART OF THIS WORK WAS SUPPORTED BY THE IST PROGRAMME OF THE EU IN THE PROJECT IST-2000-32795 SCHEMA.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

Each transversal implementation of a wavelet filter has an equivalent lifting-based implementation [9].

The most widely used lifting scheme for motion-compensated temporal transforms is the (2,2) lifting, corresponding to the classical 5/3 wavelet transform. By design, the lifting scheme implementations of motion-compensated (MCed) WT are fully invertible, regardless of the properties of the Motion Vectors (MV). Unfortunately, as shown in [10], it turns out that the MCed lifting-based (2,2) wavelet transform does not implement the equations of the original 5/3 wavelet unless the motion field $\mathbf{v}$ satisfies certain conditions, which are both invertibility and additivity. Usual MVs computation methods, based on either block matching or deformable meshes, violate these properties in the general case. Of course, it is possible to constrain them to be additive and/or invertible, but these constraints may lead to imprecise vectors, and thus, to a sub-optimal MC.

(N,0) lifting scheme [10] represent an efficient alternative to classical lifting schemes. Indeed, it has been shown that they implement exactly transversal MCed wavelet transforms while being perfectly reconstructible, provided that the MVs are precise enough. For example, the (2,0) lifting transform described below replaces the original (2,2) lifting scheme.

Let us consider a sequence $(x_n)_N$ of N images. We will further denote by $\mathbf{v}_{i+j\to i}(\mathbf{m})$ a motion vector of pixel at spatial location $\mathbf{m}$ in frame $i+j$ that displaces this pixel to the corresponding location in frame i. It follows that $x_i[\mathbf{m}+\mathbf{v}_{i+j\to i}(\mathbf{m})]$ is a motion-compensated image x_i with respect to image x_{i+j}. Using these notations, the (2,0) lifting wavelet filters with motion compensation produce the following low pass and high pass sub-bands (SBs), respectively $(l_n)_{N/2}$ and $(h_n)_{N/2}$:

$$\begin{aligned} h_k[\mathbf{m}] &= x_{2k+1}[\mathbf{m}] - \frac{1}{2}\big(x_{2k}[\mathbf{m}+\mathbf{v}_{2k+1\to 2k}(\mathbf{m})] \\ &\quad + x_{2k+2}[\mathbf{m}+\mathbf{v}_{2k+1\to 2k+2}(\mathbf{m})]\big) \\ l_k[\mathbf{m}] &= x_{2k}[\mathbf{m}] \end{aligned} \tag{1}$$

These (N,0) filters are also very convenient for temporal scalability purposes. Furthermore, they require only half of the MVs compared to regular motion-compensated lifting schemes, saving bit rate for encoding the temporal SBs themselves.

2.2. Motion estimation

The way motion is estimated plays an important role in the final performance of the coder. An ideal motion estimator should produce precise and easy-to-encode MV fields. To this end, we use a block-matching-based algorithm that minimizes a correlation criterion built on both luminance and chrominance mean square errors, rather than luminance only. We find that the chrominance information reduces the number of the particularly visible color outliers, improving the final visual quality, and smoothes the MV field, thus reducing the motion bitrate.

The use of (2,0) filters on 3 or 4 decomposition levels, combined with accurate motion estimation, leads to an efficient temporal decorrelation at a reasonable cost.

3. SPATIAL ANALYSIS AND MODEL-BASED ALLOCATION

The MCed temporal filter performs a dyadic WT on the input sequence, producing $M = L+1$ temporal SBs of the same spatial size as the original sequence (where L is the number of decomposition levels).

3.1. Problem statement

Let us assume that we have a certain encoding technique, allowing to encode the i-th SB with performances expressed by the function $D_i = D_i(R_i)$, where D_i is the distortion of the i-th SB when encoded at rate R_i. Then, optimal resource allocation problem consists in finding the rate vector $\mathbf{R} = \{R_i\}_{i=1}^M$ which minimizes the reconstruction distortion $D(\mathbf{R})$ (intended as the mean square error between original and decoded sequence) under the constraint $\sum_{i=1}^M a_i R_i \le R_{SB}$, where R_{SB} is the rate available for SBs encoding, $a_i = 2^{-l_i}$ and l_i is the level of the i^{th} SB.

In the case of orthogonal SB coding, it has been shown that D can be expressed as a sum of $D_i(R_i)$ [11]. This results has been extended to the case of biorthogonal filter [12], weighting each SB distortion term with a factor w_i which depends on the filter and on the decomposition scheme. Thus, our rate allocation problem can be written as follows:

$$D = \sum_{i=1}^{M} w_i D_i(R_i) \tag{2}$$

with the constraint $\sum_{i=1}^M a_i R_i \le R_{SB}$.

We solve this problem with a Lagrangian approach. The resulting optimal rate allocation vector $\mathbf{R}^* = \{R_i^*\}_{i=1}^M$ verifies the following set of equations:

$$\frac{\partial D_i}{\partial R_i}(R_i^*) = -\frac{\lambda}{w_i} \qquad \forall i \in \{1,\dots,M\} \tag{3}$$

where λ is the Lagrange multiplier. Thus, optimal allocation corresponds to points having the same slope on the "weighted" curves $(R_i, w_i D_i)$.

3.2. Proposed algorithm

We propose the following algorithm to find the optimal rate allocation vector. A guess value is chosen for the slope,

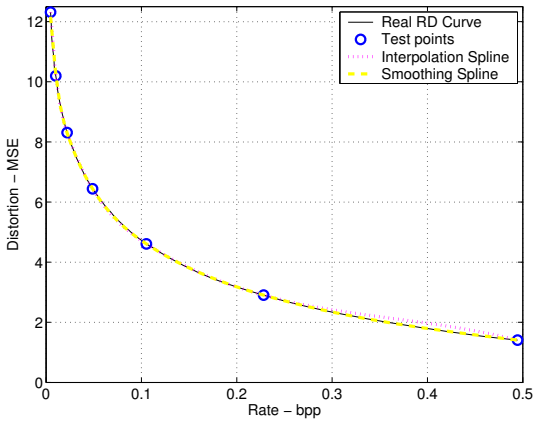

Fig. 1. Spline approximation of a real RD curve

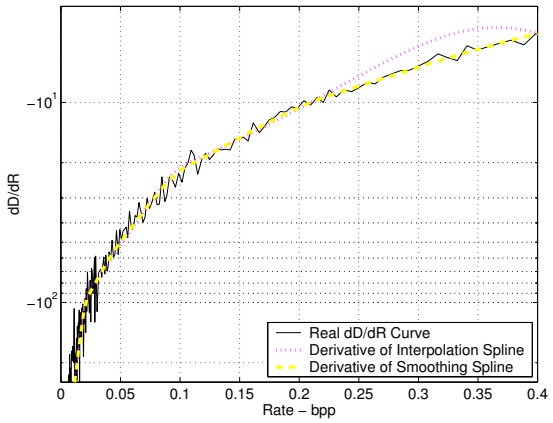

Fig. 2. RD curve: first derivative in each case

say λ. Then we look for points on the weighted RD curves where the slope is λ, by inspecting their derivatives. We consider the rates $R_i(\lambda)$ associated to these points, and sum them obtaining the corresponding total rate, say $R_t(\lambda)$. Then, if $|R_{SB} - R_t(\lambda)|$ is smaller than a given tolerance, the algorithm stops, otherwise a new value for λ is chosen with the simple bisection method.

3.3. Model for RD curves

This rate allocation algorithm requires the RD curves of SBs, or rather, their first derivative. A simple but computationally expensive way to obtain a reliable representation of these curves is to encode each SB many times at different rates and then compute resulting distortion (brute force approach). Unfortunately, since accurate estimates are needed in the whole range of possible rate allocation values, many test points are required.

To overcome this problem, we propose a model-based approach which is remarkably less complex than the regular brute force one. The main idea is to use splines (we tried both interpolation and smoothing splines [13]) as parametric model for the RD curves. From a few points (usually less than 10 is enough) of the actual curve, we can obtain the analytical representation of the spline, just by computing some parameters for each point. Then, we can compute analytically (i.e. via a few spline parameters) the first derivatives, once again with a very little computational complexity, even negligible with respect to the WT coefficients computation.

In Fig. 1, we reported, as an example, the "real" RD curve for the highest frequency SB (obtained by encoding the SB at 200 different rates), and the estimated curve computed with cubic splines using only the 7 circled points. This splines representation gives very good results, as the different curves are almost undistinguishable. For the same curve, we reported in Fig. 2 the "real" first derivative and the first derivative of splines. The proposed method completely removes irregularities from this curve. Thus, when the allocation algorithm looks for points with the same derivative, we have more robust results, especially at lower bit rates.

4. EXPERIMENTAL RESULTS

4.1. Description of the coder

We implemented the algorithm described in the previous section. As encoding technique, we chose the currently best performing one for still images, JPEG2000. Since we use 3 levels of (N,0) filters for the temporal transform, the lowest temporal SB is simply obtained by sub-sampling the input sequence, so JPEG2000 is conveniently exploited to encode it. Regarding higher temporal SBs, MC and correlation between coefficients justify the use of a still-images-oriented encoding technique, even though suitable settings are needed, like a reduced number of decomposition levels and an appropriate dynamics.

Motion is estimated using 16×16 blocks and $\frac{1}{2}$-pixel precision. We regroup MVs according to the GOP they belong to, and compose them in two images (for vertical and horizontal vector components). Then we perform a lossless JPEG2000 compression on the two obtained images. This technique, which preserves the full JPEG2000-compatibility of the coder, achieves encoding rates comparable or lower than the first order joint entropy of the MVs.

We first compared the performances of the coder with different allocation methods. To this end, we encoded the "foreman" CIF sequence, using the different methods successively, with the same number (7) of RD points. Experiments confirm that the complexity overhead due to the B-Spline interpolation of a 7-points curve is negligible with a fast implementation [14]. Resulting performances are shown in Fig. 3, where we see that the splines method outperforms the regular one by up to 0.4 dB, with smoothing spline proving to be the best. The most significant differ-

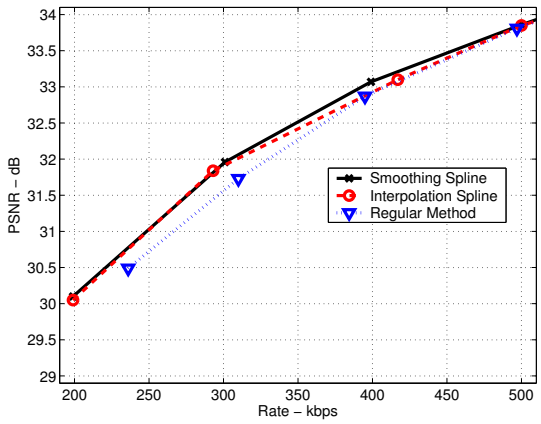

Fig. 3. Comparison of interpolation methods

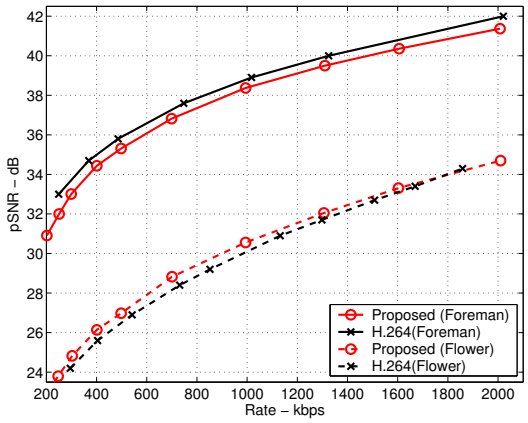

Fig. 4. Performance comparison on sequences "foreman" (solid) and "flowers and garden" (dashed) - first 64 images

ence is visible at lower rates, where the proposed algorithm takes advantage of the regularized first derivative.

Lastly, the proposed algorithm was compared with H.264 on the test sequences "flowers and garden" and "foreman", with similar motion compensation settings. Depending on the input sequence, we found (Fig. 4) that our coder achieves performances close to ("foreman" sequence) or better than ("flowers and garden" sequence) those obtained by H.264. However, whereas H.264 is not natively scalable, our coder produces a highly scalable bitstream compatible with JPEG2000. Note that the sequences obtained by our coder and analyzed in Fig. 4 have been extracted from a single bitstream.

4.2. Conclusion

In conclusion, we have presented a new video coding algorithm based on WT, fully compatible with JPEG2000, highly scalable, and matching the performances of H.264. Coding resources for different SBs are optimally allocated by means of a model-based and low-complexity algorithm. Future work will focus on better movement models and optimized rate allocation between motion vectors and SBs.

5. REFERENCES

[1] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transforms," *IEEE Trans. on Image Processing*, vol. 1, no. 2, pp. 205–220, Apr. 1992.

[2] D. Taubman and M.W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*, Kluwer Academic Publishers, 2002.

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

[4] S.J. Choi and J.W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. on Image Processing*, vol. 8, no. 2, pp. 155–167, Feb. 1999.

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

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

[7] J. Viéron, C. Guillemot, and S. Pateux, "Motion compensated 2D+t wavelet analysis for low rate fgs video compression," in *Proc. of Tyrrhenian Intern. Workshop on Digital Comm.*, Capri, Italy, Sept. 2002.

[8] C. Parisot, M. Antonini, and M. Barlaud, "3D scan based wavelet transform and quality control for video coding," *EURASIP Journal on Applied Signal Processing*, Jan. 2003.

[9] I. Daubechies and W. Sweldens, "Factoring wavelet transforms into lifting steps," *J. Fourier Anal. Appl.*, vol. 4, no. 3, pp. 245–267, 1998.

[10] T. André, M. Cagnazzo, M. Antonini, M. Barlaud, N. Božinović, and J. Konrad, "(N,0) motion-compensated lifting-based wavelet transform," in *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004.

[11] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic, Jan. 1988.

[12] B. Usevitch, "Optimal bit allocation for biorthogonal wavelet coding," in *Proc. of Data Compression Conf.*, Mar. 1996, pp. 387–395.

[13] M. Unser, "Splines: A perfect fit for signal and image processing," *IEEE Signal Processing Magazine*, vol. 16, no. 6, pp. 22–38, November 1999.

[14] F. Precioso, M. Barlaud, T. Blu, and M. Unser, "Smoothing B-spline active contour for fast and robust image and video segmentation," *IEEE Transactions on Image Processing*, 2004 (to appear).