LOSSLESS VIDEO CODING USING MULTI-FRAME MOTION COMPENSATION

Elias S. G. Carotti[1] *and Juan Carlos De Martin*[2]

[1]DAUIN/[2]IEIIT–CNR — Politecnico di Torino
c.so Duca degli Abruzzi, 24, 10129, Torino, Italy
phone: (+39) 011 564 7036, fax: (+39) 011 564 7099
email: [juancarlos.demartin|elias.carotti]@polito.it web: http://multimedia.polito.it/

ABSTRACT

In this paper we consider the problem of lossless compression of video sequences exploiting the temporal redundancy between frames. In particular, we present a technique performing motion compensation on more than one past frame. Each prediction component is optimally weighted to minimize the mean squared error of the residual. Experimental results for several standard video sequences show that multi-frame motion compensation with optimal weighting outperforms regular 1-frame motion compensation with gains up to 18.2% even for the case of just two past reference frames.

1. INTRODUCTION

Lossless compression of digital images has been the focus of many research efforts in the last decade, when many new techniques were proposed, among which CALIC [1] and LOCO-I [2] standardized as JPEG-LS [3]. Several new applications, in fact, demand compression services that do not alter the original data. Medical imaging, for instance, may require lossless compression to make sure that physicians will only analyze pristine diagnostic images [4]. Professional imaging, where images need to be stored in their original undistorted form for future processing, is another important field of application for lossless compression; also, many high-end digital cameras enable the photographer to access the raw, uncompressed picture, i.e. not altered by any coding algorithm. This can be very important if images are to be used in a production pipeline where subsequent lossy coding-decoding cycles could heavily affect the overall quality of the final results.

However, while lossless image compression has long been recognized as an important field, for what concerns video sequences, most research efforts focused on lossy coding, probably due to the wide application field involved.

Comparatively, lossless coding of video sequences has received much less attention while being increasingly important for a number of applications ranging from digital cinema, post-production, archiving and, last but not least, medical applications. For example, medical imaging applications, such as computerized axial tomography (CAT), magnetic resonance imaging (MRI) or positron emission tomography (PET), often generate sequences of strongly related images. Since a single CAT image can be as large as 130 MB, compression is clearly desirable, both for storage and remote medical applications.

Finally, one of the *digital cinema*'s main requirements is lossless video compression, due to the strict needs during acquisition, post-production, archiving and distribution. The maximum quality possible has to be preserved during all the steps in the chain from the acquisition to the film theaters [5].

Most existing lossless coding techniques are typically based on a simple paradigm consisting of a prediction step followed by context-modeling and context-based entropy coding of the residual. The aim of the prediction step is to exploit the spatial redundancy due to the regularity and smoothness of most continuous-tone images. Both CALIC and LOCO-I follow this scheme to a certain extent, first determining on a pixel basis the best predictor among a set of fixed predefined ones and then encoding the prediction residual.

Video, being a sequence of often highly correlated images, is characterized by temporal redundancy between subsequent frames, which is due to almost temporally invariant backgrounds and to objects moving across the frames. A few works have dealt specifically with this additional source of redundancy, promising higher gains with respect to independent lossless coding of each individual frame.

One of the first works dealing with video sequence was presented by Sayood *et al.* in [6] where various techniques taking into account temporal and spectral redundancy of color video sequences were presented and an adaptive scheme switching between the two sources of redundancy was proposed. CALIC, the well-known state-of-the-art algorithm for lossless still image coding, was extended to handle interframe redundancy in [7]. In [8, 9] the authors presented a low-complexity adaptive algorithm which combined, on a pixel basis, a spatial and a temporal predictor to form a prediction minimizing the MSE on a causal context of the pixel to be coded.

To accurately model motion, however, a pixel-based approach is usually not sufficient, and blocks of pixels have to be considered. Motion compensation is commonly employed to model motion of objects between subsequent frames, especially for lossy video coding standards such as MPEG and H.264 [10, 11]. Motion compensation consists in dividing each frame into small blocks and for each one of them searching a past frame (typically the preceding one) for the most similar block according to a predefined distance measure, then, the residual difference along with the relative displacement between the two blocks is coded. Thus, while motion compensation (like least-square prediction, in general) is not directly aimed at minimizing entropy, which is the ultimate goal for lossless coding techniques, it is a useful tool to obtain a lower entropy residual with respect to the original frame, because typically the prediction residual is characterized by a more peaky and skewed distribution with a lower entropy.

Motion compensation was already proven to be an effective tool for removing temporal redundancy in lossless video coding, in the above cited [6] and, more recently in [12],

Red			Green			Blue		
0.88	0.96	0.92	0.89	0.97	0.92	0.91	0.97	0.93
0.87	0.93	0.89	0.88	0.94	0.90	0.90	0.95	0.92
0.83	0.88	0.85	0.85	0.89	0.86	0.86	0.90	0.88

Table 1: Correlation coefficients between 3×3 pixel neighborhoods in the same positions in frame[i] and frame[i-1] (sequence: `mobile and calendar`).

Red			Green			Blue		
0.84	0.90	0.87	0.85	0.91	0.88	0.87	0.92	0.90
0.82	0.87	0.84	0.83	0.88	0.86	0.86	0.90	0.88
0.79	0.84	0.81	0.81	0.85	0.83	0.83	0.87	0.85

Table 2: Correlation coefficients between 3×3 pixel neighborhoods in the same positions in frame[i] and frame[i-2] (sequence: `mobile and calendar`).

where a technique combining motion estimation using the previous frame as a reference with backward-adaptive least-square prediction was proposed.

The main contributions of this paper regard motion-compensation applied to lossless coding of video sequences. In particular, we show that multiple reference frames combined with least-square weighing can considerably improve the performance of motion compensation-based lossless video coding. We experimentally prove that gains up to 18.2% can be achieved by exploiting the correlation between a frame and the two preceding, with respect to regular one frame based motion compensation. The proposed technique can be integrated into other motion compensation-based lossless video sequence compression algorithms, such as the technique described in [12].

The rest of this paper is organized as follows: in Section 2 the main sources of redundancy in a video sequence are described, the proposed algorithm is presented in Section 3, and results are discussed in Section 4; finally, conclusions are drawn in Section 5.

2. VIDEO SEQUENCE REDUNDANCY

The main sources of correlation in a color video sequence are spatial, temporal and spectral redundancy.

Spatial redundancy depends on the correlation between pixels of the same color band belonging to the same frame, and is typically very high, at least for continuous-tone natural images.

Temporal redundancy depends on the correlation between pixels of temporally adjacent frames and is typically exploited by lossy video compression techniques such as MPEG which depend on effective removal of temporal redundancy to achieve high compression ratios.

Experiments show that in several cases temporal correlation decreases slowly with time, being quite high even between frames separated by ten or more other frames. As an example, Table 1 shows the correlation coefficients for 3×3 neighborhoods in the same position in the current and the previous frames for the test video sequence `Mobile and calendar`. Table 2 shows the same information about the current frame and the frame before the previous one. Clearly, it is apparent that there is a potential gain if more than one past reference frame is used for predicting the current one.

Finally, color video sequences are characterized by another source of redundancy, which is due to the correlation between the different color bands of a frame; this is usually referred to as spectral redundancy. Typical color video sequences have three color bands (usually red, green and blue).

In this paper we address the problem of exploiting temporal redundancy as well as spatial redundancy by means of multi-frame motion compensation.

3. ALGORITHM DESCRIPTION

In this section we will briefly review regular motion compensation with one past reference frame and we propose an extension using two reference frames. When motion compensation is performed, the frame to-be-coded, i, is divided in a number of blocks of size $N \times N$; for each block $\underline{B}^i(\underline{p})$ at position $\underline{p} = (x,y)$, the previous frame $i-1$ is searched in a neighborhood of $\underline{p}$ for a block $\underline{B}^{i-1}(\underline{p}+\underline{v}) = (b_1^{i-1}, \ldots, b_{N^2}^{i-1})$ which minimizes a given distance measure; commonly employed measures are the euclidean distance between the two blocks or the sum of absolute differences.

The residual difference

$$\underline{e} = \underline{B}^i(\underline{p}) - \underline{B}^{i-1}(\underline{p}+\underline{v}), \tag{1}$$

is then entropy-coded along with the corresponding motion vector, $\underline{v}$, indicating the relative displacement of the two blocks.

We propose a scheme where more than one reference frame is used, i.e., Eq. 1 becomes

$$\hat{\underline{e}} = \underline{B}^i(\underline{p}) - \sum_{j=1}^{M} w_j \cdot \underline{B}^{i-j}(\underline{p}+\underline{v}_j), \tag{2}$$

where M is the number of past frames used for prediction and $\underline{\mathbf{W}}(\underline{p}) = (w_1(\underline{p}), \ldots, w_M(\underline{p}))$ are appropriate weights. This means that for each block $\underline{B}^i(\underline{p})$ a closest match is sought for in a number M of past frames and a prediction is formed as a weighted linear combination of the selected blocks from the preceding frames. To take into account non-stationarity across the frame, since correlation between blocks in the current frame and the past reference frames is not constant, these weights cannot be considered constant but they need to be computed for each block.

The weights $\underline{\mathbf{W}}(\underline{p})$ (for the sake of simplicity in the rest of the paper we will refer to them simply as $\underline{\mathbf{W}}$ and w_j respectively) are computed so as to minimize the Minimum Squared Error (MSE) of the residual, by solving for least-squares the system of equations:

$$\underline{\mathbf{C}} \cdot \underline{\mathbf{W}} = \underline{\mathbf{R}},$$

where

$$\underline{\mathbf{C}} = \begin{bmatrix} b_1^{i-1}(\underline{p}+\underline{v}_{i-1}) & \ldots & b_1^{i-M}(\underline{p}+\underline{v}_{i-M}) \\ \vdots & \ddots & \vdots \\ b_{N^2}^{i-1}(\underline{p}+\underline{v}_{i-1}) & \ldots & b_{N^2}^{i-M}(\underline{p}+\underline{v}_{i-M}) \end{bmatrix}$$

is a matrix whose columns are made from each prediction component block's pixels and

$$\underline{\mathbf{R}} = \begin{bmatrix} b_1^i(\underline{p}+\underline{v}_i) \\ \vdots \\ b_{N^2}^i(\underline{p}+\underline{v}_i) \end{bmatrix}$$

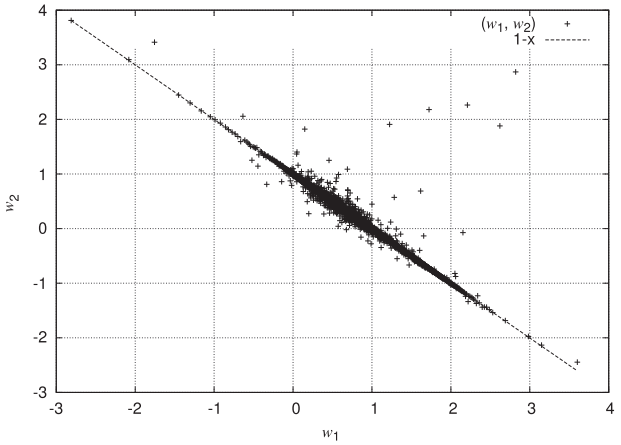

Figure 1: This pictures displays the distribution of the two weights (w_1, w_2) computed from the training set. It is evident how most of the points are on the line $w_2 = 1 - w_1$.

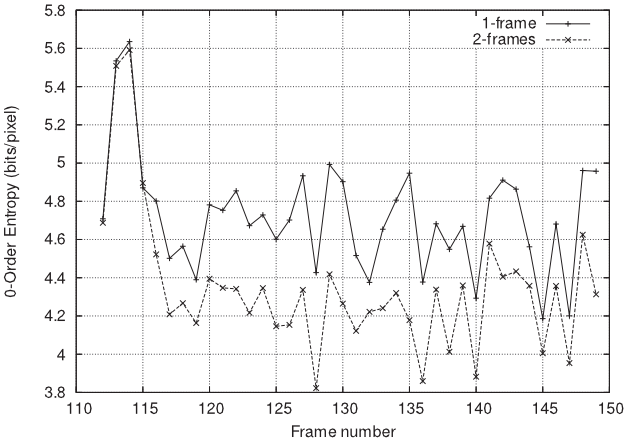

Figure 2: Zero-order Entropy for the Green band of the last 40 frames of the video sequence `Foreman` in the cases of motion compensation using one and two past reference frames. Most notably the 0-order entropy level for the two frames case is consistently under the other.

is a column vector containing the pixel values of the block to be predicted. Typically, this is an over-determined system, i.e., with more equations than unknowns, for which an exact solution cannot be found in general, but which can be can be easily and quickly solved for a solution minimizing the Mean Square Error (MSE) through SVD or QR-decomposition.

The weights $\underline{\mathbf{W}}$ for one block are not independent one from the others and, typically, their sum is about one, which is reasonable and expected because all the blocks have approximately the same energy, having been chosen to minimize the MSE with respect to $\underline{B}^i(\underline{p})$. Figure 1 shows the distribution of $\underline{\mathbf{W}}$ for the two-frame case, thus they can be quantized with a vector quantizer to minimize the bitrate needed to encode them. For this purpose, an optimal (in the MSE sense) vector quantizer can be designed on a training set and used at the encoder assuming it is known at the decoder. For each block the encoder computes the optimal weights, quantizes them, transmits the quantization index as side information and uses the corresponding quantized version $(\hat{w}_1, \ldots, \hat{w}_M)$ in Eq. 2 to compute the prediction residual, so that the decoder can invert the process and losslessly reconstruct the original frame.

On the other side, the decoder needs to be given both the motion vectors $\underline{v}_j$ and the quantization indices for $\underline{\mathbf{W}}$ so that the same prediction can be formed and added to the residual thus allowing perfect reconstruction. As a consequence, the encoded bit-stream consists of the prediction residuals and the side information, i.e., the motion vectors and the quantization indices. Of course, performing motion-compensation on M frames implies also sending M motion vectors as side information, which accounts for a slight increase in bitrate. Due to the high correlation between adjacent motion vectors, though, their entropy is very low compared to the savings achieved in coding the residuals.

Moreover, when some of the weights $\underline{\hat{\mathbf{W}}}$ tend to be approximately 1.0 for some specific frames and 0.0 for all the others, then not all the motion vectors need to be transmitted and just the relevant frames should be used for prediction, thus achieving some additional bitrate saving.

As can be easily seen the whole process is highly asymmetrical, with a fairly complex encoder and a simple decoder.

4. RESULTS

We implemented and tested the proposed technique using two past reference frames ($M = 2$) and we compared it with regular motion compensation; in both cases the test set considered was the green band of a number of standard color video sequences.

The block size was chosen to be 16×16 ($N = 16$) which is a common choice, for example in MPEG and H.264, and the search range for full motion-compensation was set to ± 8 pixels; the weights $\underline{W}$ were quantized on 5 bits, so that the amount of side information needed to transmit them could be considered negligible with respect to the gain; for the chosen block size the increase in bitrate for transmitting the weight's quantization indices is less than 0.02 bits per pixel.

Figure 2 depicts 0-order entropy for 1-frame and 2-frame motion compensation for the standard test video sequence `foreman`; it is evident how using two frames consistently delivers a lower entropy level over single-frame motion compensation.

This example is confirmed by the average results shown on Table 3 for several test sequences; for each one of them compression was performed excluding the first 50 frames, which were used to train the vector quantizer.

Gains up to 18.21% are achieved for the video sequences Salesman and Mobile & Calendar which are rich of high-frequency content. Lower gains are achieved on Akiyo and Silent, which are two similar video sequences in which a foreground person moves slowly in front of a static background; this behavior is probably due to the fact that a significant part of each picture is almost constant so that increasing the number of frames used for prediction gives a negligible contribution with respect to single frame prediction.

Sequence name	1-Frame MC bits/pixel	2-Frame MC bits/pixel	Diff. bits/pixel	Gain (%)
Salesman	3.569	4.364	0.795	18.21%
Mobile & Calendar	4.390	5.160	0.771	14.94%
Container	3.263	3.520	0.258	7.32%
Tempete	4.667	5.032	0.366	7.28%
Kitchgrass	4.315	4.632	0.316	6.83%
Sean	3.042	3.182	0.141	4.42%
Akiyo	1.905	1.970	0.065	3.30%
Silent	3.340	3.452	0.112	3.26%

Table 3: Average results for the two techniques on the test set considered in this paper; per frame average 0-order entropy values are shown. Two-frames motion compensation consistently outperforms the competitor.

5. CONCLUSIONS

In this paper we presented and discussed an extension to the basic motion compensation paradigm for lossless video coding. Motion compensation is a common building block for many video-sequence coding techniques.

After a preliminary study which proved that temporal redundancy slowly decreases with time, so that for many video sequences correlation is quite high even between frames which are separated by ten of more other frames, we proposed to use more past frames for predicting the current one and to weight each contribution to minimize the mean squared error of the prediction residual.

The proposed technique was tested on a number of standard video sequences and was proven to attain gains up to 18.2% with respect to regular single-frame motion compensation, in terms of 0-order entropy of the prediction residual, thus allowing for better packing of the data and, consequently, considerable bandwidth savings.

Future work includes the study of techniques using a higher number of past frames. Experimenting with different block and search window sizes to evaluate how they affect performance should be considered as well.

REFERENCES

[1] X. Wu and N. Memon, "Context-based, adaptive, lossless image coding," *IEEE Transactions on Communications*, 45(4):437–444, April 1997.

[2] M. J. Weinberger, G. Seroussi, and G. Sapiro, "LOCO-I: a low complexity, context-based, lossless image compression algorithm," In *Proceedings of Data Compression Conference*, pages 140–149, March 1996.

[3] ITU-T SG8, "Lossless and near-lossless compression of continuous-tone still images (ITU-T T.87—ISO/IEC 14495-1)". *ITU-T*, June 1998.

[4] I. Christoyianni, E. Dermatas, and G. Kokkinakis, "Fast detection of masses in computer-aided mammography," *IEEE Signal Processing Magazine*, 17(1):54–64, January 2000.

[5] ISO/IEC 15444-3:2002, "Information technology – JPEG 2000 image coding system – Part 3: Motion JPEG 2000," 2002.

[6] N. D. Memon and K. Sayood, "Lossless Compression of Video Sequences," *IEEE Transactions on Communications*, 44(10):1340–1345, October 1996.

[7] X. Wu, W. Choi, N. Memon, "Lossless interframe image compression via context modeling," In *Proceedings of Data Compression Conference*, pages 378–387, 1998.

[8] E. S. G. Carotti, J. C. De Martin, and A. R. Meo, "Backward-adaptive lossless compression of video sequences," In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 3417–3420, 2002.

[9] E. S. G. Carotti, J. C. De Martin, and A. R. Meo, "Low-complexity lossless video coding via adaptive spatio-temporal prediction," In *Proc. IEEE Int. Conf. on Image Processing*, volume 2, pages 197–200, September 2003.

[10] C.E. Fogg J.L. Mitchell, W.B Pennebaker and D.J. LeGall, *MPEG Video Compression Standard*. New York: Chapman and Hall, 1996.

[11] ITU-T Rec. H.264 & ISO/IEC 14496-10 AVC, Advanced video coding for generic audiovisual services. *ITU-T*, May 2003.

[12] D. Brunello, G. Calvagno, G. A. Mian, and R. Rinaldo, "Lossless compression of video using temporal information," *IEEE Transactions on Image Processing*, 12(2):132–139, February 2003.