LOW-COMPLEXITY LOSSLESS VIDEO CODING VIA ADAPTIVE SPATIO-TEMPORAL PREDICTION

Elias S. G. Carotti[1], *Juan Carlos De Martin*[2], *Angelo R. Meo*[1]

[1] Dipartimento di Automatica e Informatica/[2]IEIIT-CNR
Politecnico di Torino
Corso Duca degli Abruzzi, 24 — I-10129 Torino, Italy
E-mail: [carotti|demartin|meo]@polito.it

ABSTRACT

Lossless coding of video sequences is becoming increasingly attractive for applications as diverse as digital cinema, medical imaging and professional video processing. We present a new low-delay, low-complexity algorithm for lossless color video compression. The proposed technique adaptively exploits temporal, spatial and spectral redundancy. Key features of the coder are a backward-adaptive temporal predictor, an intra-frame spatial predictor and adaptive optimal weighting of both predictive components. The residual error is entropy coded by a context-based arithmetic encoder. The proposed lossless video encoder delivers significantly higher compression gains than traditional approaches at approximately the same complexity and delay, enabling efficient storage and communications for emerging lossless video applications.

1. INTRODUCTION

Lossless compression of digital images is becoming a mainstream technology. Several important applications, in fact, demand compression techniques that do not alter the original data. Medical imaging, for instance, often requires lossless coding to make sure that physicians will only analyze pristine diagnostic images [1]. Another emerging important area of application is professional imaging, where images need to be stored and transmitted in their original undistorted form for future processing. Even in consumer products, such as digital cameras, lossless image compression is quickly becoming a standard feature.

As storage and computational power increase, the focus is now moving to *lossless video* compression. The applications are numerous and diverse. Medical imaging applications, for instance, often generate sequences of strongly related images, as in computerized axial tomography (CAT), magnetic resonance imaging (MRI) and positron emission tomography (PET). Another application relying on lossless compression is *digital cinema* (recently the subject of an MPEG Call For Proposal), where films are to be efficiently stored and delivered in digital format while maintaining the highest possible video quality. Professional and consumer video processing systems can also increasingly benefit from efficient lossless video compression techniques.

Lossless image compression (mostly addressing gray scale images) has received a considerable amount of interest in recent years (e.g., [2] [3]). Lossless coding of video sequences, instead, has received significantly less attention.

A hybrid lossless video compression approach exploiting temporal, spatial and spectral redundancy was investigated in [4]. More recently, an inter-band version of the lossless image coding algorithm CALIC was proposed in [5]. A low-complexity, backward-adaptive technique for compression of color video sequences, was proposed in [6]. The algorithm was based on a fixed combination of two predictors to remove spatial as well as temporal redundancy. The high non-stationarity of video sequences, however, was not adequately addressed by such time- and content-invariant approach.

We propose a spatio-temporal lossless compression technique which matches the instantaneous characteristics of the input video signal. The contributions of the spatial and the temporal predictors are, in fact, adaptively weighted with pixel-level granularity to locally minimize the energy of the prediction residual. Temporal prediction is pixel-based, achieving good performance at a fraction of the complexity of the block-based motion estimation approach. A simple yet effective technique for removing spectral redundancy is also adopted. The residual error is entropy-encoded with a context-based arithmetic coder.

This paper is organized as follows. In Section 2 the results of a preliminary investigation of the main kinds of redundancy that characterize video sequences are presented. In Section 3, the proposed lossless coding algorithm is described. Test results are presented in Section 4. Finally, conclusions are discussed in Section 5.

2. VIDEO SEQUENCES REDUNDANCY

The main sources of correlation in a video sequence are *spatial*, *temporal* and *spectral* redundancy. Spatial redundancy depends on the correlation between same color-band pixels belonging to the same frame; it is typically high, at least for continuous-tone natural images. Many algorithms exist in literature to effectively remove spatial redundancy and some are used for reversible compression of gray-scale still images; among them, LOCO-I [7], standardized as JPEG-LS [8], and CALIC [9].

Temporal redundancy is due to the correlation between pixels of temporally adjacent frames. Lossy video compression techniques such as MPEG depend heavily on effective removal of temporal redundancy to achieve high compression ratios. Experiments showed that temporal redundancy decreases slowly with time, being still significant on average for many sequences even between

This work was supported in part by CERCOM, the Center for Wireless Multimedia Communications, Torino, Italy, http://www.cercom.polito.it.

0-7803-7750-8/03/$17.00 ©2003 IEEE

frames separated by ten or more other frames.

Finally, another source of redundancy is due to the correlation between the color bands of a multi-spectral image. Typical video sequences have three color bands (usually red, green and blue). The proposed technique includes a differential encoding scheme that effectively removes part of the spectral redundancy of color video sequences.

3. CODER DESCRIPTION

The proposed compression algorithm consists of two main steps: adaptive signal prediction followed by context-based arithmetic coding of the prediction residual. The predictive step aims to decorrelate the frames of the video sequence. Two separate predictors are used: a novel mixed spatial-temporal predictor, working separately on each color band, and a purely spatial predictor. An adaptive optimal weight to combine the two predictions is computed and the resulting error is then corrected for local prediction biases. Spectral decorrelation and entropy coding follow.

In the remaining part of this paper the generic i-th frame of a video sequence is referenced as frame[i]. The pixel at location (x, y) of frame[i] is indicated by $p_i(x, y)$.

3.1. Temporal Prediction

The temporal predictor predicts the color of each pixel based on the values of those pixels located in the same pixel neighborhood in the previous two frames. This backward-adaptive approach is simple and effective, and has the advantage of not requiring the transmission of any side information.

If frame[i] is the current frame, then the proposed temporal predictor is a function of pixel values in frame[$i-1$] and frame[$i-2$]. For each pixel $p_i(x, y)$ in each color band of frame[i], in fact, we look at the corresponding pixel in frame[$i-1$], referred to as the *reference pixel*. We then analyze the pixel in the same position in frame[$i-2$] as well as its eight neighbors, computing the nine differences between them and the reference pixel. The differences are saved in the 9×9 matrix,

$$\Delta = \begin{pmatrix} \Delta_{-1,-1} & \Delta_{0,-1} & \Delta_{+1,-1} \\ \Delta_{-1,0} & \Delta_{0,0} & \Delta_{+1,0} \\ \Delta_{-1,+1} & \Delta_{0,+1} & \Delta_{+1,+1} \end{pmatrix},$$

where $\Delta_{j,k} = |p_{i-1}(x, y) - p_{i-2}(x + j, y + k)|$.

The smallest matrix term is taken as an indication that the corresponding pixel of frame[$i-1$] is the best predictor of $p_i(x, y)$. The implicit assumption for this choice is that the motion registered between frame[$i-2$] and frame[$i-1$] continues at the present time.

To model sub-pixel motion the prediction can be further refined using as predictor a linear combination of the pixels corresponding to the smallest matrix terms. The average value of matrix Δ, Δ_{avg}, is computed. Prediction is then based on those pixels for which the corresponding matrix terms (whose sum is Δ_{sum}) are lower than Δ_{avg} (an offset is added before the computation to deal with null terms). The prediction corresponds to a weighted average of those pixels using the following weights :

$$w_{j,k} = \frac{\Delta_{jk}}{\Delta_{sum}}.$$

The final predicted value is, therefore:

$$\dot{p}_i(x, y) = \sum_{j,k|\Delta_{j,k} \leq \Delta_{avg}} w_{j,k} \cdot p_{i-1}(x + j, y + k) \quad (1)$$

with $j, k \in \{-1, 0, +1\}$.

3.2. Spatial prediction

To take into account spatial redundancy, a prediction based on the values of pixels belonging to the same frame of the pixel to be coded is computed. We employed the well-known LOCO-I spatial predictor used in JPEG-LS, the so-called Median Edge Detector (MED)[7]. This spatial predictor estimates the pixel to be coded based on the values of the three already coded neighboring pixels.

The spatially predicted value is:

$$\hat{p}_i(x, y) = \begin{cases} \min(a, b), & \text{if } c \geq \max(a, b) \\ \max(a, b), & \text{if } c \leq \min(a, b) \\ a + b - c, & \text{otherwise,} \end{cases} \quad (2)$$

where $a = p_i(x, y-1), b = p_i(x-1, y), c = p_i(x-1, y-1)$.

This predictor is characterized by simple edge-detection capabilities as well as minimal delay and complexity.

3.3. Spatio–Temporal Prediction

The final predicted value is given by the linear combination of the temporal and spatial predictor, i.e.:

$$\tilde{p}_i(x, y) = \alpha \cdot \dot{p}_i(x, y) + (1 - \alpha) \cdot \hat{p}_i(x, y), \quad (3)$$

where $\dot{p}_i(x, y)$ and $\hat{p}_i(x, y)$ are the temporally and spatially predicted values, respectively, and α is an averaging weight, which can be either fixed or adaptive.

3.3.1. Optimal fixed weighting

The simplest choice for α is to set it to a reasonable constant value, i.e., 0.5, for all sequences; this was the approach used in [6] and is justified by the observation that, on average, temporal and spatial redundancy are approximately equally strong. Video material, however, can be quite diverse. Figure 1 shows the mean square prediction error as a function of α for the test video sequence *News*; the performance varies greatly depending on the value of α. For each video sequence there will be a specific optimal value of α that will maximize the overall spatio-temporal prediction gain.

The objective is thus to minimize the mean square error of the final prediction error signal, i.e.:

$$\min_{\alpha} MSE_P = \mathrm{E}\{(p_i(x, y) - \tilde{p}_i(x, y))^2\}. \quad (4)$$

Deriving and minimizing with respect to α, the optimal value for a given sequence, α_{opt}, is obtained:

$$\alpha_{opt} = \frac{\mathrm{E}\{(p_i(x, y) - \hat{p}_i(x, y)) \cdot (\dot{p}_i(x, y) - \hat{p}_i(x, y))\}}{\mathrm{E}\{(\dot{p}_i(x, y) - \hat{p}_i(x, y))^2\}}. \quad (5)$$

The value α_{opt} is the optimum relative weight for the two predictors, i.e., the weight which leads to the best decorrelation results for a given sequence.

Table 1 compares the compression performance obtained using the video-sequence-dependent optimal weight α_{opt} with the

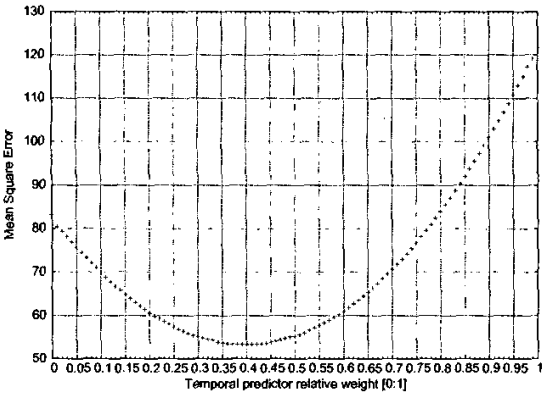

Fig. 1. Mean square error of the prediction residual signal as a function of α, for the video sequence *News*.

sequence-invariant fixed weight approach ($\alpha = 0.5$). The gain is significant in practically all cases, reaching 16.4% for the *Amélie trailer* sequence.

Computation of the optimal weight, however, is possible only if the whole sequence is available. Moreover, it requires a significant amount of computation and memory.

3.3.2. Optimal adaptive weighting

The computation of α_{opt} requires infinite delay and is computationally expensive since it requires two coding passes, one to evaluate α_{opt} and the second to actually code the sequence. Moreover, video sequences are non-stationary signals: a locally optimum weight for each color pixel should therefore significantly outperform long-term optimal weighting.

The locally optimum weight, $\alpha_N(x,y)$, as in Equation (4), can be obtained by minimizing the energy of the residual signal over a rectangular window of N pixels centered around the pixel to be coded. To avoid sending side information to the receiver a backward-adaptive approach is adopted in this case too. The window therefore encompasses the most recent N past samples.

Windows of various sizes were tested; results showed that a small window of just 12 past samples (corresponding to the last 4 RGB color pixels) delivers nearly optimal performance. Moreover, $\alpha_N(x,y)$ can be easily computed in an incremental way with a pixel-by-pixel update. Determining the locally optimum $\alpha_N(x,y)$ is, therefore, characterized by limited memory and computation requirements.

3.4. Bias cancellation

A context-based bias cancellation step follows the prediction process, to limit the negative effect of local biases of the predictor. Bias cancellation is also another way to make the prediction adaptive, since it adds a corrective term depending on the context.

For each pixel a causal context is identified. The residual error

$$\tilde{e} = p_i(x,y) - \tilde{p}_i(x,y)$$

is then corrected with the average of the errors already made in that context, e_{avg}, obtaining a new residual error $\hat{e} = \tilde{e} - e_{avg}$.

Video sequence name	Compressed file size for fixed α (kilobytes)	Compressed file size for optimum α (kilobytes)	Gain
Amélie trailer	251,640	210,352	16.4%
Bus	24,592	22,300	9.3%
Calcio	62,404	55,480	11.1%
Container	18,896	18,356	2.9%
Foreman	21,256	19,816	6.8%
IT	90,748	85,252	6.1%
Mobile and C.	44,392	44,192	0.5%

Table 1. Compression gain using the sequence-optimal weight α_{opt} with respect to the fixed weighting case ($\alpha = 0.5$).

The causal context of the generic pixel $p_i(x,y)$ is defined as three gradients, (g_1, g_2, g_3), computed as the differences between the pixel values at positions $(x-1, y-1)$, $(x-1, y)$, and $(x, y-1)$ in the current frame, and the predicted pixel value $\tilde{p}_i(x,y)$. Since this information is known to the decoder, perfect reconstruction is possible without the need of side information. These gradients have a double-sided exponential probability mass function (p.m.f.), centered at zero, and are therefore non-linearly quantized on 16 levels, forming the triplet $Q = (q_1, q_2, q_3)$, that defines the context. After bias-cancellation is performed, the context is updated with $\tilde{e}$ so that it reflects the error just made by the predictor without the correction.

To take into account signal non-stationarity, the accumulated error values for each context are periodically halved so that the average error follows the behavior of the predictor.

The residual error $\hat{e}$, belongs to the range:

$$-\tilde{p}_i(x,y) \leq \hat{e} \leq N - \tilde{p}_i(x,y),$$

where N is the alphabet size minus one (typically 255 for 24-bit per pixel RGB color frames, where each color band is coded with 8 bits). Since the predicted value $\tilde{p}_i(x,y)$ is known to the decoder, the sign-bit can be discarded. By this simple symbol-mapping scheme, we obtain a new residual error $\dot{e}$ such that $0 \leq \dot{e} \leq N$. This operation does not alter the error distribution, but prevents the error-range to become too large when subsequent prediction schemes are applied in cascade. Moreover the adaptive entropy coder can thus work with a reduced symbol-set, attaining higher compression ratios.

3.5. Spectral decorrelation

After the last prediction step, spectral redundancy is even higher than the original, and the residual images have a smooth gray-scale look. We opted to employ a differential coding scheme, where two of the colors are represented by the differences with respect to the third color. Experience has shown that, for RGB-color frames, it is better to code $\dot{e}_r$ and $\dot{e}_b$ as

$$\begin{cases} \ddot{e}_r = \dot{e}_g - \dot{e}_r, \\ \ddot{e}_b = \dot{e}_g - \dot{e}_b. \end{cases} \quad (6)$$

Again the result is mapped to the interval $[0, N]$ by removing the sign bit.

		JPEG-LS	Adaptive Spatio-Temporal Compression		
Video sequence name	Uncompressed file size (kilobytes)	Compressed file size (kilobytes)	Compressed file size (kilobytes)	Compression gain with respect to JPEG-LS	Compression ratio w.r.t. the original
Amélie trailer	764,416	222,660	197,540	11.3%	3.9
Bus	44,992	27,628	22,332	19.2%	2.0
Calcio	119,560	59,304	55,604	6.2%	2.2
Container	44,992	23,680	18,356	22.5%	2.5
Foreman	44,992	23,440	19,704	15.9%	2.3
IT	233,100	88,080	80,668	8.4%	2.9
Mobile and Calendar	75,392	52,736	44,012	16.5%	1.7
News	44,992	20,144	15,064	25.2%	3

Table 2. *Compression performance of the proposed adaptive spatio-temporal lossless video compression technique.*

3.6. Entropy coding

The difference frame is then entropy coded with a context-driven arithmetic coder. As usual the process is assumed to be generated by a number of different sources variously interlaced, with different variances. Context modeling is an effective way to deal with such kind of sources. The contexts are determined by scalar quantization of the same gradients already computed for bias cancellation. However, not to incur into the well-known context dilution problem, only two of the three gradients and a different quantizer are used, so that they are mapped on a lower number of bits; in our experiments we used three bits, for a total of 64 different contexts. Moreover, assuming that the p.m.f.'s of the gradients are symmetrical, only the absolute part is quantized and then used to drive the arithmetic coder.

4. RESULTS

We tested the proposed technique with a number of standard test video sequences. Due to the lack of well-known lossless video compression algorithms, we considered as a reference a video coding scheme consisting of lossless image coding of each frame of a sequence; such approach has been used in practice for several lossless video applications. The state-of-the-art JPEG-LS lossless image coding standard, based on LOCO-I, was used [8].

Table 2 describes the performance of the proposed technique. The file sizes corresponding to the JPEG-LS frame-by-frame coding algorithm are reported in column 3, while the file sizes for the proposed spatio-temporal compression technique are shown in column 4. The last but one column shows the compression gain of the proposed technique with respect to JPEG-LS; it can be seen that it consistently outperforms JPEG-LS by up to 25.2% in terms of compressed file sizes.

Compression ratios with respect to the uncompressed file sizes are also reported in the last column. Ratios range between 2 and 4 for a wide selection of video material, confirming the compression potential of the proposed lossless video coding approach.

5. CONCLUSIONS

We presented a new technique for lossless coding of video sequences. The coding scheme effectively removes temporal, spatial and spectral redundancy, thus allowing an adaptive entropy encoder to attain high compression ratios. Key features of the coder are a backward-adaptive spatio-temporal predictor, an intra-frame spatial predictor and an adaptive optimal combination of both.

Compared to traditional approaches to lossless video coding, the proposed algorithm delivers significantly higher compression with low delay and limited complexity.

With respect to the original uncompressed material, bandwidth and storage can now sustain between 2 and 4 times more video data, confirming the potential of lossless video coding techniques for practical applications.

6. REFERENCES

[1] I. Christoyianni, E. Dermatas, and G. Kokkinakis, "Fast detection of masses in computer-aided mammography," *IEEE Signal Processing Magazine*, vol. 17, no. 1, pp. 54–64, January 2000.

[2] R. B. Arps and T. K. Truong, "Comparison of international standards for lossless still image compression," in *Proceedings of the IEEE*, June 1994, vol. 82, pp. 889–899.

[3] B. Carpentieri, M. J. Weinberger, and G. Seroussi, "Lossless compression of continuous-tone images," in *Proceedings of the IEEE*, November 2000, vol. 88, pp. 1797–1809.

[4] N. D. Memon and K. Sayood, "Lossless Compression of Video Sequences," *IEEE Transactions on Communications*, vol. 44, no. 10, pp. 1340–1345, October 1996.

[5] X. Wu, W. Choi, N. Memon, "Lossless Interframe Image Compression via Context Modeling," in *Proceedings of Data Compression Conference*, 1998, pp. 378–387.

[6] 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*, 2002, pp. 3417–3420.

[7] 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*, March 1996, pp. 140–149.

[8] 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.

[9] X. Wu and N. Memon, "Context-based, adaptive, lossless image coding," *IEEE Transactions on Communications*, vol. 45, no. 4, pp. 437–444, April 1997.