POLYPHASE SPATIAL SUBSAMPLING MULTIPLE DESCRIPTION CODING OF VIDEO STREAMS WITH H264

*R. Bernardini, M. Durigon, R. Rinaldo**

Università degli Studi di Udine - Italy
Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica
Via delle Scienze 208, Udine, e-mail: rinaldo@uniud.it

*L. Celetto, A. Vitali**

ST Microelectronics
Via C. Olivetti n. 2, 20041, Agrate Brianza - Italy
email: {luca.celetto, andrea.vitali}@st.com

ABSTRACT

In this work, we propose a Multiple Description (MD) coding system for video streams. In particular, our scheme originates four descriptions from the spatially downsampled polyphase components of the original frames. Each description is compressed independently with the recent H264/AVC video coding standard, it is packetized and sent over an error prone network. In case of errors in one or more descriptions, appropriate concealing is applied at the receiver, before insertion of the corrected frames into the corresponding receiver frame buffers. We propose and compare different concealment solutions and a post processing stage to attenuate visual effects related to MD coding. We analyze the trade off between robustness to channel errors and coding efficiency, comparing the proposed technique with Single Description (SD) video coding with H264/AVC. Experimental results validate the effectiveness of the proposed scheme.

1. INTRODUCTION

In error prone environments, the transmission of multimedia material like audio, video or still images, is subject to a set of constraints which traditional coding and transmission systems are usually not designed for. This is particularly true when the communication system has real-time constraints or when there is no backward channel. If losses are inevitable, representations that make all the received packets useful can be of great benefit. MD coding techniques [1] are designed to perform this goal by creating several descriptions of the original source. If any description is lost or corrupted by channel failures, the other descriptions can be used to reconstruct the original signal with acceptable quality. The larger the number of the correctly received descriptions is, the higher the quality of the reconstructed information source. The emerging application of real time video transmission over unreliable networks and channels, e.g., in wireless systems, seems to be the natural environment for the application of MD coding techniques. In this paper, we propose an MD video coding system which originates four descriptions from the spatially downsampled polyphase components of the original frames. Thus, each description has dimension 1/4 of the original frame, it is compressed independently with the recent H264/AVC video coding standard, it is packetized and sent over an error prone network. If packets from one description are lost, the corresponding corrupted spatial region is recovered from the correctly received descriptions. Moreover, the restored frame is inserted in the frame buffer of the corresponding decoder to mitigate the effect of error propagation in motion compensated differential decoding. In Section 2 we provide an overview of the proposed MD video coding system, underlying advantages and problems of such an approach. In Section 3 we propose and analyze the performance of different concealment techniques, which use spatial redundancy among descriptions to recover from channel errors. In Section 4 we analyze an intrinsic problem of MD video coding that can introduce visual degradation of the reproduced video material and propose a post processing step to mitigate this type of artifact. In Section 5 we compare the performance of the proposed MD system with traditional SD coding.

*This work has been supported under project FIRB-PRIMO (Reconfigurable Platforms for Wideband Wireless Communications) of the Italian Ministery of University and Research , MIUR.

2. OVERVIEW OF THE SYSTEM

The proposed scheme increases robustness to channel errors by creating different descriptions from the original video stream. According to general MD coding requirements, equal importance is given to each description and the scheme is not strictly related to a specific video coding algorithm. In particular, the input video sequence is processed frame by frame, and the spatial polyphase components are extracted from the original frame. This originates four substreams with spatial dimension one fourth of the original. The four descriptions are then processed by independent H264 / AVC coders [3],[4], but any other coder could be used.

On the receiver side there are four synchronised H264 / AVC decoders. They simultaneously process the four subframes corresponding to one original frame. This allows for coherent spatial error concealment, and is not a limitation especially for streaming or real time applications, where high delay is not tolerated. Decoders are connected to a restoring block which reassembles descriptions at the original full size. The restoring block performs error concealment if some packets are lost (Section 3), and applies a post filtering operation to attenuate MD coding artifacts (Section 4). Spatial polyphase decomposition is then applied to the recovered full size frame, and each component is copied into the corresponding decoder frame buffer. This prevents error propagation from reference frames due to interframe coding. Note that any video codec other than H264 / AVC could be used, since the proposed scheme simply requires read/write access to frame buffers at the decoders.

3. ERROR CONCEALMENT IN THE RESTORING BLOCK

Depending on the coder strategy, packets for each description contain data which are necessary to reconstruct spatial regions, or slices, in the corresponding subframe. Lost packets result in missing spatial polyphase components in corresponding regions of the full size reassemled frame. Pixel error patterns may have different configurations, based on the number of descriptions lost and the slice partitioning solutions adopted by the encoder. The five basic loss patterns are shown in Figure 1.

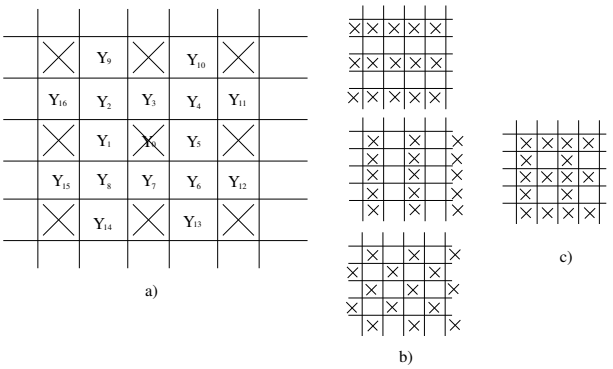

Fig. 1. Luminance loss patterns for a) one description lost, b) two descriptions lost, c) three descriptions lost.

The proposed MD coding system performs error concealment by using the high spatial correlation among descriptions, the purpose being the interpolation of missing pixels from the received ones. In the following, we propose and compare different interpolation strategies.

Besides classical interpolation algorithms such as Near Neighbour Replication (NNR), Bilinear Interpolation and Bicubic Interpolation we also consider non-linear interpolators, and propose two schemes inspired by the predictors used for lossless image coding [5] and Bayer mask demosaicing with CCD camera sensors [6]. Classical linear interpolators tend to perform a low pass operation to reconstruct lost pixels, and this is visible as a sort of smoothing effect and artifacts around natural edges. This degradation adds to the effects of coding inefficiency of MD systems that we will consider in Section 4. On the other hand, non linear interpolators try to reproduce missing pixels by replicating natural edges. We consider only algorithms with complexity comparable to those of classical linear interpolators. Pixel predictors, like the GAP predictor of the lossless image coding algorithm CALIC and the MED predictor of JPEG-LS, usually process image pixels in scan-raster order, to allow the decoder to perform the same operations. Demosaicing algorithms, instead, cope with interpolation patterns more similar to those we consider in our application. Both techniques evaluate a set of gradients in the neighborhood of the current pixel, and replicate the edge structure recognized by gradient analysis.

Below, we describe in some detail the proposed algorithms for MD error concealment and compare their performance.

3.1. Linear Interpolators

We consider two types of linear interpolators: the NNR and the bilinear algorithm. NNR replicates the first correctly received pixel in the 8-pixel neighborhood of the current one, starting from the left and proceeding in clockwise order. The bilinear interpolator reconstructs the missing pixels by averaging between adjacent pixels. Depending on the loss pattern of Figure 1, it can use four pixels, or only two pixels. For example, for one description lost, pixel Y_0 is restored by averaging Y_1, Y_3, Y_5 and Y_7.

3.2. Non Linear Interpolators

In the following, we will consider the use of non linear interpolators only when a single description is lost, i.e., for pattern a) in Figure 1. With the other loss patterns, gradient computation involves pixels too far from the current one, and non linear interpolation gives no significative improvement over the bilinear method. We propose two solutions, that we will refer to as the Edge Sensing (ES) Algorithm and the Variable Number of Gradients (VNG) Algorithm.

3.2.1. Edge Sensing

The algorithm is inspired by the predictors used for lossless image coding [5]. The algorithm aims at detecting horizontal and vertical edges around the processed pixel, and compute missing pixels taking the edge orientation into account. Two gradients (ΔH and ΔV) are computed as $\Delta H = |Y_1 - Y_5|$ and $\Delta V = |Y_3 - Y_7|$, and the interpolated value is computed as

if ($\Delta H < T$ *and* $\Delta V > T$)
$Y_0 = (Y_1 + Y_5)/2$
else if ($\Delta H > T$ *and* $\Delta V < T$)
$Y_0 = (Y_3 + Y_7)/2$
else
$Y_0 = (Y_1 + Y_3 + Y_5 + Y_7)/4$

The first condition detects an horizontal edge, the second one a vertical edge, otherwise the usual bilinear interpolation is used. The algorithm is not very sensitive to the value of T and we experimentally found that for a large set of 8 bit/pixel test natural images a threshold value of about 50 gives the best performance. The computational complexity of ES is roughly the same of that of bilinear interpolation.

3.2.2. Variable Number of Gradients

The algorithm is inspired by the one presented in [7], with the difference that it considers only one color component and gradient computation is modified accordingly. As it will be seen, its performance is similar to that of ES, at the expense of increased complexity. A set of 8 gradients is computed from the 16 luminance pixel neighborhood of Y_0, shown in Figure 1.a,

$G_1 = 2|Y_1 - Y_5| + 0.5\big(|Y_3 - Y_{16}| + |Y_2 - Y_3| + |Y_7 - Y_8| + |Y_7 - Y_{15}|\big)$
$G_2 = 2|Y_2 - Y_6| + |Y_3 - Y_9| + |Y_1 - Y_{16}|$
$G_3 = 2|Y_3 - Y_7| + 0.5\big(|Y_1 - Y_2| + |Y_1 - Y_9| + |Y_4 - Y_5| + |Y_5 - Y_{10}|\big)$
$G_4 = 2|Y_4 - Y_8| + |Y_3 - Y_{10}| + |Y_5 - Y_{11}|$
$G_5 = 2|Y_1 - Y_5| + 0.5\big(|Y_3 - Y_4| + |Y_3 - Y_{11}| + |Y_6 - Y_7| + |Y_7 - Y_{12}|\big)$
$G_6 = 2|Y_2 - Y_6| + |Y_5 - Y_{12}| + |Y_7 - Y_{13}|$
$G_7 = 2|Y_3 - Y_7| + 0.5\big(|Y_1 - Y_8| + |Y_1 - Y_{14}| + |Y_5 - Y_6| + |Y_5 - Y_{13}|\big)$
$G_8 = 2|Y_4 - Y_8| + |Y_1 - Y_{15}| + |Y_7 - Y_{14}|$

Each gradient corresponds to a different direction ($G_1 \leftrightarrow$West, $G_2 \leftrightarrow$ North West,...). A threshold value is computed as $T =$

$k_1\text{Min} + k_2(\text{Max} - \text{Min})$, where $k_1 = 1.5$ and $k_2 = 0.5$, Min and Max being the minimum and maximum values of the set of computed gradients. The subset I of gradients with absolute value less than T is selected. One pixel is associated to every gradient, $G_i \leftrightarrow Y_i$, and the missing pixel Y_0 is obtained by averaging pixels $Y_i, i \in I$. The idea is to locate those pixels that are likely to be similar to the pixel under consideration and compute the missing pixel value accordingly.

Figure 2 compares the PSNR of the considered algorithms for some 8 bit/pixel test images, when the first polyphase component of each image is reconstructed from the others. Average values

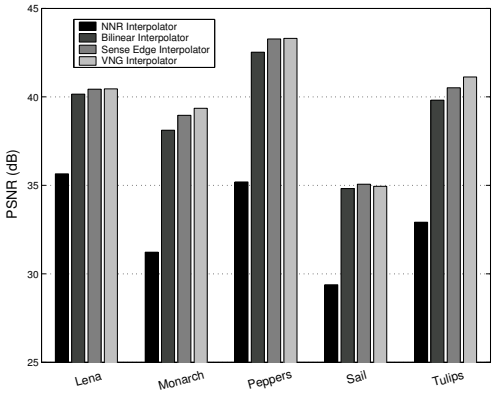

Fig. 2. PSNR for a set of 8 bit/pixel images; one description lost.

are reported in Table 1. It is evident that bilinear and non linear adaptive interpolators have the best performance, with ES representing the best compromise between performance and computational complexity. Edge-directed non linear interpolators also appear preferable from a subjective quality point of view.

Table 1. Average PSNR for a set of 8 bit/pixel images, one description lost.

NNR	Bilinear	Edge Sensing	VNG
32.87 dB	39.09 dB	39.65 dB	39.84 dB

It is useful to consider the performance of the algorithms for coded video frames, to take into account the effect of coding quantization error. Figure 3 plots the reconstruction quality (PSNR) for the sequence *Foreman* in CIF format, coded with H264 / AVC at different rates (QP=10 ÷ 40). As before, the first polyphase component in each decoded frame is reconstructed from the others. It can be noted that at low rates (high QP), the performance difference between bilinear, Edge Sensing and VNG becomes negligible, due to substantial coding smoothing effect.

4. POST FILTERING IN THE RESTORATION BLOCK

One problem with the proposed MD video coder is a spatial granularity artifact in the reconstructed sequence at medium-low rates, due to the fact that the four polyphase component streams are coded independently (see the image on the left in Figure 4). To reduce the effect, we propose an adaptive post filtering operation, which removes the artifact from flat regions in the reconstructed frames while preserving natural edges. It is crucially important to

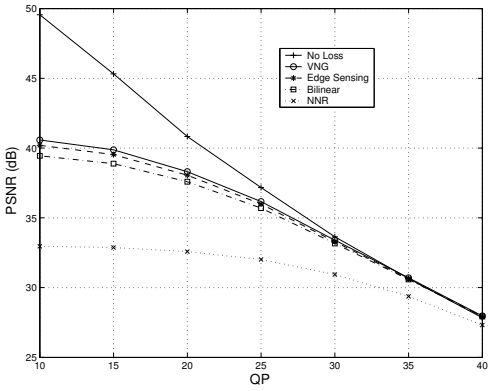

Fig. 3. PSNR for the CIF Sequence Foreman, compressed with H264 / AVC; one description lost.

be able to distinguish between true edges in the image from discontinuities due to the granularity effect created by independent coding of the single descriptions. To preserve image sharpness, the true edges should be left unfiltered as much as possible.

The separable 2D extension of a simple noncausal three tap filter is applied to the rows and columns of each frame. The filtered value $\hat{x}_k$ is calculated as

$$\hat{x}_k = a \cdot x_{k-1} + (1-2a) \cdot x_k + a \cdot x_{k+1}, \tag{1}$$

where x_k are the pixel values of the full size reconstructed frames. The filter is a low-pass linear phase filter, with a parametrizing the filter band. The value $a = 0.25$ is chosen to have a multiplication free filtering operation. The value of $\hat{x}_k$ is substituted to x_k only if $|x_k - x_{k-1}| < \beta$ and $|x_k - x_{k+1}| < \beta$. This condition is used to evaluate spatial activity, and limit the filtering operation to flat regions. The threshold β is determined by taking into account the quantization process of the video coding algorithm. The idea is to perform filtering only if the difference between consecutive pixels is larger than a value implied by the coder quantization process. H264 / AVC DCT coefficient quantization can be assumed to be uniform, with a quantization step Δ that varies exponentially with the quantization parameter QP [9]

$$\Delta = \Delta_0(2^{QP/6} - 1) \tag{2}$$

Errors on reconstructed frames are related to quantization via the inverse DCT. Given $x = s + e$, where s is the original signal, x the reconstructed one and e the error on reconstructed samples, it is easy to show that a simple bound on this error is given by

$$|e| \leq G\Delta_0(2^{QP/6} - 1) \triangleq E \tag{3}$$

where G is the constant sum of absolute IDCT matrix row coefficients. By assuming that the error e has a normal distribution, its variance σ^2 can be related to E by relation $E = \gamma\sigma$. Threshold β can be estimated such that $P(|e| \leq \beta) = 1/100$, resulting in $\beta/\sigma \simeq 1.3$. Therefore, a suitable choice for β is given by

$$\beta = \frac{1.3E}{\gamma} = \frac{1.3G\Delta_0(2^{QP/6} - 1)}{\gamma} = \beta_0(2^{QP/6} - 1) \tag{4}$$

By chosing $\gamma = 6.5$, equation (4) becomes

$$\beta = 0.5(2^{QP/6} - 1) \tag{5}$$

Simulation results on several test video sequences with different choices of β, show a very good agreement between the value of β that maximises performance, and the one computed with equation (5). In the experiments, if $\beta < 6$, filtering is actually turned off.

The PSNR gain due to post filtering is shown in Table 2 for the CIF sequence Foreman, and for some values of QP. For relatively high QP (medium to low bit rate) the objective gain is about 0.6 dB. The results are relative to the transmission of the four polyphase components with no loss.

Table 2. Increase in PSNR due to Post Filtering operation for the Foreman sequence coded with H264 / AVC.

	QP = 24	QP = 30	QP = 36
MD	37.69 dB	33.64 dB	30.09 dB
MD & Post Filtering	38.12 dB	34.25 dB	30.69 dB

To appreciate the visual quality improvement, a detail of one frame before and after the application of the post filtering procedure is shown in Figure 4.

Fig. 4. Detail of one frame before (left) and after (right) the application of the Post Filtering procedure.

5. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed MD video coder, where channel errors in each description are corrected using the ES algorithm described in Section 3, and the post filtering procedure of Section 4 is applied in the restoring block. We compare the proposed solution with a standard Single Description H264 / AVC coder which includes basic error concealment as described in [10]. Both solutions are based on the H264 / AVC test model software version *JM6.0a*. To increase robustness to channel errors and make a fair comparison, the SD coder uses the Random Intra Macro Block Refresh coding option, i.e., 100 Macro Blocks for every CIF frame are coded in intra-mode. The MD coder utilizes four independent H264 / AVC coders for the QCIF substreams corresponding to the spatial polyphase components, but no Random Intra Macro Block Refresh coding option is activated. The other coding options are the same for the SD and MD coders. In particular, the GOP structure is IBBBBPBBBBPBBBBPBBBBI and the slices have a fixed 400 byte dimension. Each slice is sent as a packet, and each packet is lost independently with probability P_l. Results are averages of 50 independent transmission trials. Figure 5 shows the rate-distortion comparison between the SD and MD coders for various values of P_l for the CIF luminance video sequence *Foreman*. The sequence is 100 frame long. From the figure, it can be seen that the MD system performs better than the SD one as soon as $P_l > 0.01$. For $P_l = 0.01$, the SD PSNR curve is slightly above that of the MD coder. Also in this case, however, the artifacts introduced by error concealment in the reference SD coder are often very annoying, whereas MD concealment typically produces no visual quality loss in the reconstructed video stream.

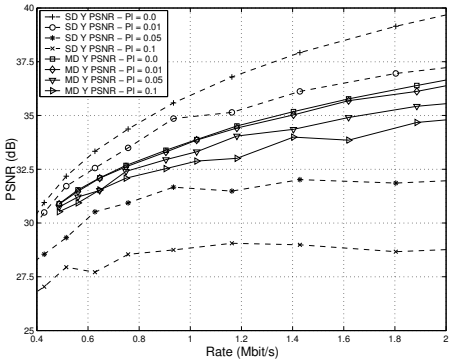

Fig. 5. Comparison between MD and SD coding systems, for the CIF sequence Foreman compressed with H264 / AVC.

6. REFERENCES

[1] V. K. Goyal, "Multiple Description Coding: Compression Meets the Network", *IEEE Signal Proc. Mag.,* Sept. 2001, pp. 74-93.

[2] V. K. Goyal, J. Kovacevic, "Generalized Multiple Description Coding with Correlating Transforms", *IEEE Transactions on Information Theory,* VOL. 47, NO. 6, Sept. 2001

[3] Joint Video Team of ITU-T and ISO/IEC JTC 1, ITU-T Rec. H.264 — ISO/IEC 14496-10 AVC, March 2003.

[4] T. Wiegand, G. J. Sullivan, G. Bjontegaard, A. Luthra, "Overview of the H.264 / AVC Video Coding Standard", *IEEE Transactions on Circuit and Systems for Video Technology,* VOL. 13, NO. 7, July 2003

[5] N. Memon, X. Wu, "Recent Devolpements in Context-Based Predictive Techniques for Lossless Image Compression", *The Computer Journal,* Vol. 40, no. 2/3, 1997

[6] R. Ramanath, W. E. Snyder, "Adaptive demosaicking", *Journal of Electronic Imaging* 12(4), 633-642, Oct. 2003

[7] E. Chang, S. Cheung, D. Pan, "Color Filter Array Recovery Using a Threshold-based Variable Number of Gradients", *Proc. IS&T/SPIE Int. Symp. on Electronic Imaging: Science and Technology,* Volume 3650: San Jose, Jan. 1999

[8] P. List, A. Joch, J. Lainema, G. Bjøntegaard, M. Karczewicz, "Adaptive Deblocking Filter", *IEEE Transactions on Circuit and Systems for Video Technology,* VOL. 13, NO. 7, July 2003

[9] H. S. Malvar, A. Hallapuro, M. Karczewicz, L. Kerofsky, "Low-Complexity Transform and Quantization in H.264 / AVC", *IEEE Transactions on Circuit and Systems for Video Technology,* VOL. 13, NO. 7, July 2003

[10] T. Stockhammer, M. M. Hannuksela, T. Wiegand, "H.264/AVC in Wireless Environments", *IEEE Transactions on Circuit and Systems for Video Technology,* VOL. 13, NO. 7, July 2003