CONTENT ADAPTIVE WYNER-ZIV VIDEO CODING DRIVEN BY MOTION ACTIVITY

João Ascenso[1], Catarina Brites[2], Fernando Pereira[3]

[1] joao.ascenso@lx.it.pt, [2] catarina.brites@lx.it.pt, [3] fp@lx.it.pt
[1] Instituto Superior de Engenharia de Lisboa – Instituto de Telecomunicações
[2,3] Instituto Superior Técnico – Instituto de Telecomunicações

ABSTRACT

In distributed video coding (DVC), the video statistics are exploited, partially or totally at the decoder. A particular case of DVC, Wyner-Ziv video coding deals with lossy source coding with side information at the decoder and allows moving part or the entire motion estimation task to the decoder. In this context, it is the decoder responsibility to obtain the side information, a guess of the encoded Wyner-Ziv frame and the encoder only sends parity bits to improve its quality. In this paper, a technique targeting the improvement of the quality of the side information, and thus of the rate-distortion performance of the Wyner-Ziv codec is proposed. This is achieved by adaptively adjusting the size of the motion interpolation structure (or GOP length) according to the motion activity along the sequence. Experimentally, this allows to achieve gains up to 0.8 dB without performing any motion estimation or complex mode decision at the encoder.

Index Terms — Video coding, motion analysis

1. INTRODUCTION

Today's hybrid video coding schemes tackle temporal redundancy by encoding the difference between the motion compensated predicted frame (obtained from past and future decoded frames) and the current frame. Then the DCT transform and entropy coding tools are applied to the residual error and a complex bitstream is assembled and sent to the decoder.

For distributed video coding architectures such as the one adopted here [1, 2], the encoder assumes that side information, which may be understood as a guess of the actual frame, is available at the decoder and sends parity bits in order to improve its quality. The higher the correlation between the side information and the current frame, the better is the guess and thus the quality of the side information and the fewer are the parity bits to be sent from the encoder to the decoder to achieve a target quality.

In conventional hybrid video coding, this is achieved by using as predictor the motion-compensated predicted frame, both at the encoder and at the decoder. It is thus the task of the encoder to perform the complex process of motion estimation. For the distributed video coding solutions known in the literature, this task has been moved totally (or at least partially) to the decoder. The decoder estimates the current frame by looking at neighboring frames, establishing the motion trajectories between them by motion estimation and then inferring the side information by motion compensation. This biggest challenge here is to take the best advantage of existing temporal correlation by using efficient motion estimation and compensation tools at the decoder.

If this task is to predict a forthcoming frame at the decoder based on past frames, it is called frame extrapolation. If this task is to predict a frame between two temporal adjacent frames is called frame interpolation. The focus of this paper is in the latter case; however, for low-delay solutions, motion extrapolation techniques are needed.

One common approach is to perform frame interpolation using successive groups of a fixed number of pictures, which is normally referred as a closed GOP (Group of Pictures) solution. However, it would be an advantage to be able to use varying GOP sizes, exploiting better the temporal correlation of the frames inside the GOP by adapting better the temporal interpolation structure to the content. To the Wyner-Ziv codecs considered here implies developing an intelligent way to control the insertion of keyframes to separate the WZ coded frames. This paper proposes a novel content adaptive Wyner-Ziv video codec which depending on the sequence characteristics, is able to dynamically adapt the temporal coding structure through the control of the GOP size, to better exploit the video statistics and achieve a better RD performance. This new approach requires an efficient but simple GOP size control mechanism at the encoder, and a powerful and flexible frame interpolation mechanism at the decoder able to work with GOPs of any length. The GOP size control mechanism to be added to the encoder shall not significantly increase its complexity, i.e. shall not perform any motion estimation or complex mode decision. It is also proposed a block-based motion-compensated frame interpolation algorithm responsible for the generation of the side information and able to handle fast motion and long GOP sizes. This framework is essential to guarantee an efficient performance when long GOP sizes are selected by the encoder.

2. IST WYNER-ZIV VIDEO CODEC

The novel IST-WZ solution proposed in this paper is an evolution of the Wyner-Ziv coding architecture proposed in [1, 2] by the same authors. Figure 1 illustrates the new overall coding architecture.

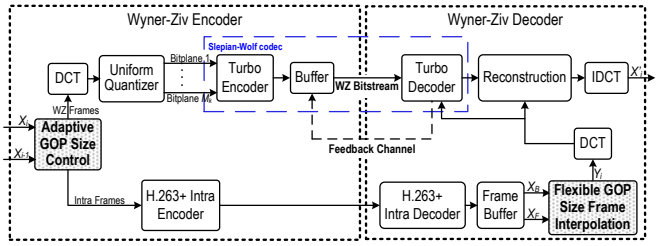

Figure 1 – *Architecture of the IST-WZ codec.*

Regarding the previous architecture two major modules are proposed to add content adaptation capabilities to the codec: 1) an adaptive GOP size control mechanism at the encoder and 2) a flexible GOP size frame interpolation framework at the decoder.

First, the coding process divides the video frames into keyframes and Wyner-Ziv (WZ) frames. The keyframes are the first frames sent from the encoder to the decoder for each GOP and are encoded using a conventional Intra coding scheme; in this case the low complexity

The work presented here was developed within DISCOVER, a European Project (http://www.discoverdvc.org), funded under the European Commission IST FP6 programme.

Intra mode of the H.263+ coding standard is used. The WZ frames are interpolated with a predefined order and the encoder use parity bits to improve its quality.

To dynamically divide the frames into keyframes and WZ frames, the novel GOP size control mechanism is added, abandoning the rigid GOP size approach used in the literature. At the encoder, the novel shaded module performs the proposed content adaptive GOP size control based on the current frame X_i and the previous frame X_{i-1} and classifies each frame as WZ or keyframe. The WZ frame is coded by applying a 4×4 block-based discrete cosine transform (DCT) as defined by the H.264/AVC standard. Then, the DCT coefficients of the entire frame X_i are grouped together, according to the position occupied by each DCT coefficient within the 4×4 blocks, forming the DCT coefficients bands. Each band is uniform quantized and bitplanes are formed and sent to the turbo encoder. The turbo coding procedure for the DCT coefficients band starts with the most significant bitplane and generates the respective parity sequences which are stored in the buffer and transmitted in small amounts upon decoder request via the feedback channel, which is available in many applications. The exploitation of the feedback channel for rate control purposes when it exists is not a limitation but rather the efficient exploitation of an existing capability.

At the decoder, the new shaded flexible GOP size frame interpolation module generates the side information Y_i, an estimate of the X_i frame, based on two references, one temporally in the past and another in the future. The Laplacian distribution is used as the correlation model between Y_i and X_i for each DCT coefficient [2]. The iterative turbo decoder uses the received parity bits and the side information to generate the decoded (almost error free) quantized symbol stream. The side information is also used in the reconstruction module, together with the decoded quantized symbol stream, to help in the X_i reconstruction task [2].

3. ADAPTIVE GOP SIZE CONTROL MECHANISM BASED ON HIERARCHICAL CLUSTERING

The objective of the novel mechanism proposed in this Section is to exploit better the temporal redundancy in the video by adapting the temporal structure of the coding process. In this case it is used the GOP length, in order to increase the RD performance, notably regarding more rigid and not content adaptive approaches. In traditional video coding a related problem exists: find the best GOP size and frame coding arrangement in terms of allocation of I, P and B frames in a dynamic way to obtain an improved rate-distortion performance [3]. Normally the algorithms performing this task make use of the motion information, typically available at the encoder, to perform complex motion analysis, to detect scene changes and to position the various types of frames inside the GOP. However, in Wyner-Ziv video coding is difficult to implement this type of techniques, especially because detailed motion information, i.e. the motion vectors or the residual frame are not available to avoid increasing the complexity of the encoder; low encoding complexity is one of the DVC main goals. Given these constraints, it is necessary to develop efficient and low complexity techniques able to decide on the GOP size by measuring the activity along the video sequence but without requiring the computational extensive motion estimation process.

3.1. Motion Activity Metrics

The simple but powerful metrics proposed in this paper to evaluate the activity along the video sequence make use of low level features, also used for video parsing and indexing [4] of large video databases. It is proposed to combine the following four features: 1) Difference of Histograms (DH); 2) Histogram of difference (HD); 3) Block histogram difference (BHD) and 4) Block variance difference (BVD). They are defined as:

$$DH(i,j)=\frac{1}{D_f}\sum_{k=0}^{L}\left|h_i(k)-h_j(k)\right| \quad (1)$$

$$HD(i,j)=\frac{1}{D_f}\left(\sum_{k=0}^{L/2-\alpha}h_{i-j}(k)+\sum_{k=L/2+\alpha}^{L}h_{i-j}(k)\right) \quad (2)$$

$$BHD(i,j)=\sum_{b=0}^{D_B/D_f}\sum_{k=0}^{L}\left|h_i(b,k)-h_j(b,k)\right| \quad (3)$$

$$BVD(i,j)=\sum_{b=0}^{D_B/D_f}\sum_{k=0}^{L}\left|\sigma_i^2(b,k)-\sigma_j^2(b,k)\right| \quad (4)$$

where i, j represent the frame index, h is the histogram operator (luminance only) with L levels, D_f and D_B are the frame and block size respectively and σ^2 the variance. For the metric HD, α is the threshold that represents the closeness to the origin. Based on experiments, the best performance is achieved for L = 32 for DH, L L = 64 for HD, L = 8 for BHD, D_B = 8×8 and α = 16.

The first two metrics work at the frame level and detect changes in global motion, e.g. zooming, panning and scene changes. The HD metric is quite effective since when significant changes occur (e.g. due to high motion) between frames, more pixels are distributed away from the origin, which causes a high value of HD. The BHD and BVD metrics are more sensitive to local motion and overcome some of the problems of the frame level metrics, e.g. object motion in static background.

3.2. Hierarchical Clustering Based on Motion Activity

The encoder must perform the GOP length selection, depending on the motion activity in the sequence. Intuitively, when the amount of motion is high, correlation is low and smaller GOP sizes must be chosen. When the amount of motion is low, correlation is higher (and thus more efficiently explored within the decoder) and longer GOPs sizes must be chosen. This strategy avoids the penalty of inserting an Intra coded frame when is not needed, which normally accounts in a significant decrease in efficiency when compared to a well interpolated WZ frame. To perform the GOP length decision, this paper proposes a new hierarchical clustering algorithm to temporally segment the sequence. For a given maximum GOP size of M frames, typically determined by the random access requirements, if they exist:

1. Calculate the four metrics above for adjacent frames and construct the 4th dimensional vector, for N_c = M-1 frames with $i > 0$:

$$x_i=\left[DH(i-1,i),\quad HD(i-1,i),\quad BHD(i-1,i),\quad BVD(i-1,i)\right] \quad (5)$$

2. Normalize the vector according to:

$$x_i=N_c x_i \Big/ \sum_{j=0}^{N_c}x_j \quad (6)$$

3. Accumulate the motion of each set of frames with the motion of the previous set of frames:

$$y_i=x_{i-1}+x_i,\quad 1<i\le N_c \quad (7)$$

4. Find the index c of the minimum accumulated motion value:

$$c=\arg\min_i\left\|y_i\right\|,\quad 1<i\le N_c \quad (8)$$

5. Cluster the motion of the corresponding frames according to:

$$x_{i-1}=\begin{cases}x_i, i>c\\ y_i, i=c\end{cases},\quad c\le i\le N_c \quad (9)$$

6. Set N_c= N_c -1

7. Go back to step 3 until the following stop criteria is satisfied:

$$y_c > \varphi \quad \vee \quad N_c = 1 \tag{10}$$

In each iteration of the algorithm (steps 2-6), frames with similar motion content are grouped, by constructing GOPs which accumulated less motion (and thus are better correlated) hierarchically. The threshold φ enables to control the maximum amount of correlation allowed and it should be adjusted according to the reliability of the frame interpolation framework, i.e. it should be adjusted according to the efficiency of the frame interpolation algorithm available at the decoder. The maximum GOP size M is constrained by memory and delay requirements, since it is necessary to store M frames at the encoder to perform the GOP length decision.

4. FLEXIBLE GOP SIZE FRAME INTERPOLATION FRAMEWORK

A frame interpolation algorithm is used at the decoder in order to generate the side information. The choice of the technique used can significantly influence the IST-WZ codec RD performance. When the correlation between the side information and the frame to be encoded (Y_i is more similar to X_i) is high, the fewer are the parity bits that need to be requested from the encoder to reach a certain quality. Another important issue to consider in the motion interpolation framework is the capability to work with longer GOPs without a significant decrease in the quality of the interpolated frame, especially when the correlation of the frames in the GOP is high. This is a complex task since the interpolation and quantization errors are propagated inside the GOP, when the frame interpolation algorithm uses as references WZ decoded frames. In this context, it is proposed in this paper a new frame interpolation framework which extends the work in [1] for GOPs of any length including longer and high motion GOPs. Figure 2 shows the architecture proposed for the novel frame interpolation scheme.

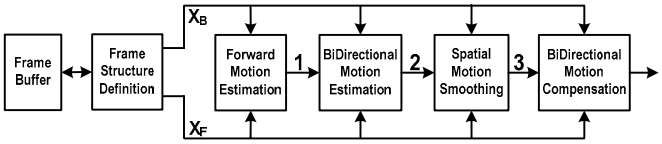

Figure 2 – *Proposed frame interpolation framework.*

In the following, the various modules in the proposed frame interpolation framework are described in detail. Its main novelty, not available in the literature, is the capability to create efficient side information for distributed video coding using arbitrary length GOPs which allow to improve the RD performance through better adaptation to the video activity behavior.

4.1. Frame Interpolation Structure Definition

The frame interpolation structure used to generate the side information is based on previously decoded frames, X_B and X_F, the backward (in the past) and forward references (in the future). For a GOP length of 2, X_B and X_F are the previous and the next temporally adjacent Intra coded frames. For other GOP lengths, the proposed frame interpolation structure definition algorithm is based on [5]; however [5] only works for GOP lengths powers of 2. The novel algorithm works for any GOP length as follows:

1. Choose the longest interframe distance N_i between two already decoded frames X_B and X_F.

2. Interpolate the frame X_i in half of the time interval according to:

$$X_i = \left\lfloor \left(X_B + X_F \right) / 2 \right\rfloor \tag{11}$$

3. Goto step 1 until all WZ frames are decoded.

Figure 3 illustrates this concept for a GOP size of 5; the numbers indicate the decoding order. Similar frame interpolation structures are used for longer GOP lengths.

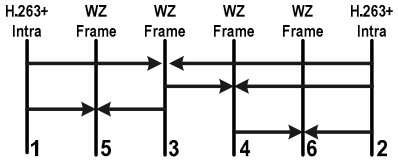

Figure 3 – *Frame interpolation structure.*

4.2. Forward Motion Estimation

A block matching algorithm is used to estimate the motion between the decoded frames X_B and X_F. In order to reduce the number of motion vector outliers, X_B and X_F are low-pass filtered first. One important issue in the design of the FME algorithm is the matching criteria used when motion estimation is performed. Generally, the goal is to minimize a cost function *CF* that measures the prediction error for a given block, e.g. the mean absolute difference (MAD) for the luminance block:

$$(v_x, v_y) = \arg\min_{d_x, d_y} CF(d_x, d_y), d_x = [-M, M], d_y = [-M, M] \tag{12}$$

$$MAD(d_x, d_y) = \frac{1}{N} \sum_{(x,y)\in B} \left| X_F(x,y) - X_B(x+d_x, y+d_y) \right| \tag{13}$$

where M is the search range, N is the number of pixels in the block, B is the block of X_F where is looked for the "best match" (v_x,v_y), according to the *CF* criteria, in frame X_B. The candidate motion vectors must be contained within the search range; in this case with integer pixel precision. However, using (12) and (13) to minimize the individual block distortion results in a quite noisy motion vector field, especially when the search range is large. To perform frame interpolation efficiently is necessary to obtain a motion vector field close to the real motion and the minimization of the frame distortion is not a good approach [1]. One possible alternative proposed here, is to add a penalizing term the popular MAD criteria and define *CF* in (12) according to:

$$CF(d_x, d_y) = MAD(d_x, d_y) \times \left(1 + K \times \sqrt{d_x^2 + d_y^2}\right) \tag{14}$$

where K is a smoothness constant that controls how much penalty is introduced when the motion vectors go to extreme positions of the search range (experimental results suggest K = 0.05). This criteria allows to increase the search range of 8 used in [1] to 32 since it regularizes the motion vector field by favouring motion vectors closer to the origin. The increase in search range enables to increase the quality of the interpolated frame when high motion occurs or longer GOPs are selected. After obtaining the motion field between X_B and X_F, for each non-overlapped block of the interpolated frame it is selected the motion vector that intersects closer to its center (see [1] for details).

4.3. Bidirectional Motion Estimation

The next step is to refine the motion vectors obtained in the previous step by using a bidirectional motion estimation scheme. The novel algorithm proposed here consists in the following steps:

1. Set the block size to N×N and for each block:

2.1 Set the search range in the backward and forward reference frames based on the motion vectors of neighboring blocks:

$$\begin{aligned} x_U + N \le d_x \le x_B - N \\ y_L + N \le d_y \le y_R - N \end{aligned} \tag{15}$$

Figure 4 shows the positions of x_U, x_B, y_L, y_R in frame X_F and illustrates the search range selected based on the neighboring motion vectors.

2.2 Find motion vector (v_x,v_y) which minimizes (12) and (14) using:

$$MAD(d_x, d_y) = \frac{1}{N} \sum_{(x,y)\in B} \left| X_F(x - N_1 d_x, y - N_1 d_y) - X_B(x + N_2 d_x, y + N_2 d_y) \right| \quad (16)$$

where N_1 and N_2 corresponds to the temporal distances to the X_F and X_B reference frames and the motion vector (dx, dy) was the one obtained in the previous module and scaled accordingly. This formula enables to select a linear motion trajectory between X_F and X_B passing at the center of the blocks in the interpolated frame.

3. Set the block size to N/2×N/2, using as initial estimate the output of step 2.2 and repeat steps 2.1 and 2.2 for each block. For sequences at QCIF resolution as used in the experiments performed later N =16.

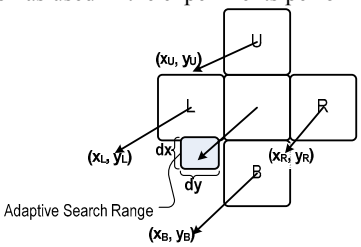

Figure 4 – *Search range adaptation.*

This technique combines a hierarchical block size technique with a new adaptive search range strategy in a very efficient way. The hierarchical coarse-to-fine approach tracks fast motion and handles large GOP sizes in the first iteration and then achieves more detail by using smaller block sizes. The new adaptive search range takes advantage of the high correlation of the true motion field to enable the correction of discontinuities, by including information from its neighbors. It performs better than a fixed search range, since it restricts the trajectory of the motion vector to be inline with the true motion of the region while still leaving room to increase its accuracy and precision.

4.4. Spatial Motion Smoothing and Bidirectional MC

Next, a spatial motion smoothing algorithm is used to make the final motion vector field smoother, except at object boundaries and uncovered regions [1]. Then, the interpolated frame can be filled by using bidirectional motion compensation as defined in standard video coding schemes. Since the time interval between the reference X_B and the interpolated frame can be different to the time interval between the interpolated frame and the reference X_F, linear interpolation weighted by the corresponding time differences to the reference frames is applied.

5. EXPERIMENTAL RESULTS

To evaluate the coding efficiency of the proposed scheme, several experiments have been performed. First, the motion-compensated frame interpolation algorithm is evaluated as a standalone tool. This indicates the quality of the side information used by the Wyner-Ziv decoder, which influences significantly the overall RD performance. Table 1 shows the average PSNR (in dB) results for the interpolated frame computed over the whole sequence.

Table 1 – *Frame interpolation performance*.

			Frame interpolation steps			
Sequence	**AVG**	**[1]**	**1**	**2a**	**2b**	**3**
Coastguard	26.32	30.36	29.65	30.32	30.39	30.59
Foreman	27.02	28.72	28.54	28.65	28.67	28.85
News	31.42	32.65	32.64	32.74	32.78	32.80
Stefan	19.47	20.65	21.42	21.40	21.39	21.53

The results are compared against a simple average frame interpolator (AVG) and the scheme in [1]. It is also shown the results obtained by applying the motion field as estimated after the n^{th} step of the algorithm shown in Figure 2. The steps 2a and 2b correspond to the 1st and 2nd iteration of the Bidirectional ME module. It is used a GOP size of 4 with lossless keyframes in order to evaluate only the contribution of the motion interpolation errors. As observed the algorithm is able to perform better than simple averaging and [1]; with gains up to 4.3 dB and 0.9 dB, respectively.

Next, it is presented the overall coding performance for the pixel-domain (IST-PDWZ) and transform-domain (IST-TDWZ) Wyner-Ziv codecs [1, 2] using the proposed adaptive GOP size control and frame interpolation framework. It is compared against the rigid GOP scheme and H.263+ Intra. The number of Intra frames is made equal in the fixed and adaptive schemes, by setting the threshold φ in the adaptive scheme in order to match the number of Intra frames in the rigid approach. The Intra frames quantization parameter QP was set in order to have approximately constant quality in the whole reconstructed video sequence.

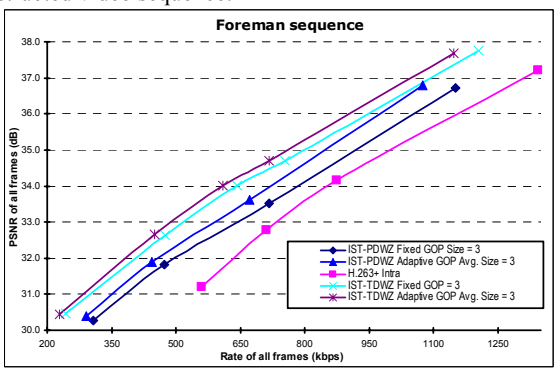

Figure 5 – *RD performance of the proposed techniques.*

Experimental results show that exploiting the temporal correlation adaptively can improve the performance up to 0.8 dB for the pixel domain codec and up to 0.4 dB for the transform domain-approach when compared to the rigid GOP size approach. It is also possible to observe gains up to 2.5 dB when compared to H.263+ Intra codec.

6. FINAL REMARKS

The main contribution of this paper is to present an algorithm to perform content adaptive GOP size control for Wyner-Ziv video coding according to the amount of motion present in the sequence. An efficient motion interpolation framework was also proposed to handle high motion and large GOP sizes. As future work, it is planned to test the contribution of each metric to the final RD performance for a wide range of sequences; looking forward to eliminate some computation without coding efficiency loss.

7. REFERENCES

[1] J. Ascenso, C. Brites, and F. Pereira, "Improving Frame Interpolation with Spatial Motion Smoothing for Pixel Domain Distributed Video Coding", *5th EURASIP Conf. on Speech, Image Processing, Multimedia Communications and Services*, Smolenice, Slovak Republic, July 2005.

[2] C. Brites, J. Ascenso, and F. Pereira, "Improving Transform Domain Wyner-Ziv Video Coding Performance", *IEEE International Conf. on Acoustics, Speech and Signal Processing*, Toulouse, France, May 2006.

[3] A. Y. Lan, A. G. Nguyen, and J.-N. Hwang, "Scene-Context-Dependent Reference-Frame Placement for MPEG Video Coding", *IEEE Transactions on CSVT*, Vol. 9, No. 3, April 1999.

[4] J. Lee and B.W. Dickinson, "Multiresolution Video Indexing for Subband Coded Video Databases", *Proc. IS&T/SPIE, Conf. on Storage and Retrieval for Image and Video Databases*, San Jose, California, USA, February 1994.

[5] A. Aaron, E. Setton, and B. Girod, "Towards Practical Wyner-Ziv Coding of Video", *IEEE International Conf. on Image Processing*, Barcelona, Spain, September 2003.