Wyner-Ziv Stereo Video Coding using a Side Information Fusion Approach

José Diogo Areia, Catarina Brites and Fernando Pereira
Instituto Superior Técnico – Instituto de Telecomunicações
Lisbon – Portugal

João Ascenso
Instituto Superior de Engenharia de Lisboa – Instituto de Telecomunicações
Lisbon – Portugal

***Abstract* — Wyner-Ziv coding, also known as distributed video coding, is currently a very hot research topic in video coding due to the new opportunities it opens. This paper applies the distributed video coding principles to stereo video coding, to propose a practical solution for Wyner-Ziv stereo coding based on mask-based fusion of temporal and spatial side informations. The architecture includes a low-complexity encoder and avoids any communication between the cameras/encoders. While the rate-distortion (RD) performance strongly depends on the motion-based frame interpolation (MBFI) and disparity-based frame estimation (DBFE) solutions, first results show that the proposed approach is promising and there are still issues to address.**

Keywords - Wyner-Ziv coding, distributed video coding, transform domain, stereo video

I. INTRODUCTION

Nowadays, most video coding solutions are monoview, where only one camera captures the video sequence. In recent years, the interest in multiview systems has been increasing, where several cameras capture the same scene from different viewing points. This scenario enables new and interesting approaches to video coding and can turn out to be one of the key technologies for a wide variety of future interactive multimedia applications. However, while some applications can still follow the conventional predictive coding approach for multiview coding with all views jointly coded, other applications like wireless surveillance, sensor networks, etc. ask for an architecture where the various cameras do not communicate among them and it is left to the decoder to exploit the inter-view correlation. To address this requirement, distributed video coding is an interesting solution since its foundation theorems, Slepian-Wolf [1] and Wyner-Ziv [2], state that separate encoding with joint decoding of two statistically dependent sources does not induce any compression efficiency loss when compared to the joint encoding used in the traditional predictive coding paradigm.

Reviewing the literature, the first conclusion is that the exploitation of distributed source coding principles in a multiview setup is a relatively new research area, especially for stereo setups. In [3], inter-view correlation can be exploited for an N-cameras setup by disparity compensation from neighbouring cameras at the decoder; in this case, the temporal side information is always disparity compensated. The experiments performed showed that the disparity-compensated side information reduced the bitrate by up to 10% over decoding without (disparity-compensated temporal side information) side information; without disparity compensation, the gain decreases to 3%. In [4], a three-camera multiview scenario is addressed where a central view is WZ coded based on two available views. There are two types of side information which are fused at the frame level: temporal side information and homographic side information which relates the side views and the central one to be WZ coded. Reported gains are around 0.5 dB at low bit rates and 0.2 dB at high bit rates with respect to simple temporal side information.

This paper proposes a pixel-level side information creation approach fusing temporal and inter-view side information through a binary mask generated at the decoder. The paper is organized as follows: Section II briefly presents the codec architecture; Sections III, IV and V detail the algorithms for the creation of temporal, inter-view and fused side information with special emphasis in terms of novelty for Section V. Finally, Section VI presents and analyses the experimental results while Section VII concludes the paper.

II. THE WYNER-ZIV VIDEO CODECS

The transform domain Wyner-Ziv (TDWZ) monoview video codec [5] which is used in this paper to create a stereo video codec is an evolution of the pixel domain Wyner-Ziv (PDWZ) codec [6]; both codecs were developed at Instituto Superior Técnico (IST). This codec is an evolution of the one initially proposed in [7] and uses a feedback channel based turbo coding approach. The TDWZ coding architecture works as follows: a video sequence is divided into Wyner-Ziv (WZ) frames and key frames. The key frames may be inserted periodically with a certain Group of Pictures (GOP) size or an adaptive GOP size selection process may be used depending on the amount of temporal correlation in the video sequence [8]; most results available in the literature use a GOP of 2 which means that odd and even frames are key frames and WZ frames, respectively. While the key frames are traditionally intraframe coded, the WZ frames are DCT transformed, quantized and turbo coded and the parity bits are stored in the buffer and transmitted in small amounts upon decoder request via the

[1] The work presented was developed within VISNET II, European Networks of Excellence (http://www.visnet-noe.org).

feedback channel (see 2nd view in Figure 1). At the decoder, the frame interpolation module is used to generate the side information frame, an estimate of the WZ frame, based on previously decoded frames, X'_B and X'_F. For a Group Of Pictures (GOP) length of 2, X_B and X_F are the previous and the next temporally adjacent key frames. The side information (SI) is then used by an iterative turbo decoder to obtain the decoded quantized symbol stream. The decoder requests for more parity bits from the encoder via the feedback channel whenever the adopted request stopping criteria has not been fulfilled; otherwise, the bitplane turbo decoding task is considered successful. The side information is also used in the reconstruction module, together with the decoded quantized symbol stream, to help in the WZ frame reconstruction task. After all DCT coefficients bands are reconstructed, a block-based 4×4 inverse discrete cosine transform (IDCT) is performed and the reconstructed WZ frame is obtained. To finally get the decoded video sequence, decoded key frames and WZ frames are mixed conveniently.

In a stereo video coding scenario, there are two dependent views to be coded. In a Wyner-Ziv coding framework, the target is to exploit in the best way the available statistical dependency not only in time, as was done for the monoview case, but also in space, i.e. between the two dependent views. In this paper, the stereo video coding scenario adopted assumes that the first view is already coded, e.g. in a conventional way using H.264/AVC, and distributed video coding principles are to be applied to the coding of the second, dependent view. In this context, the WZ monoview architecture in [5] has to be changed in order to exploit not only the temporal correlation but also the inter-view correlation as shown in Figure 1.

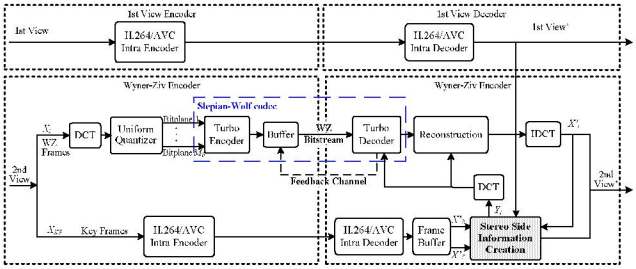


Figure 1 – Stereo video codec architecture.

The main novelty of this paper regards the development of efficient solutions for the Stereo Side Information Creation module in the new architecture. It is well known that temporal redundancy in video data is typically very high; this is the main reason for the excellent compression performance obtained with conventional video codecs. At the same time, it is also well know that inter-view redundancy is typically not as high as temporal redundancy, while strongly depending on the content, e.g. scene geometry, calibration and illumination issues. Still, in stereo video coding, the issue is about exploiting both temporal and inter-view correlations in an efficient way. While conventional predictive solutions, such as the Multiview Video Coding (MVC) standard under development by MPEG, try to exploit all redundancy at the joint encoder, WZ stereo video coding tries to do this at the decoder.

III. Creating Temporal Side Information

In order to exploit the temporal redundancy both in monoview and stereo video, the WZ video decoder has to temporally estimate the side information for each WZ coded frame through motion compensated frame interpolation. The choice of the technique used can significantly influence the WZ codec RD performance. The monoview frame interpolation framework included in the TDWZ codec used in this paper accepts GOPs of any length and is mainly based on the solution proposed in [8]. Figure 2 shows the architecture of the adopted motion-based frame interpolation (MBFI) scheme.

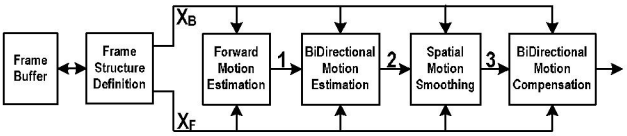


Figure 2 – Motion-based frame interpolation architecture.

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 (in the future) references. Due to its importance for the inter-view side information, the Forward Motion Estimation (FME) module is briefly explained. In this module, 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. To estimate the motion, a cost function for the block prediction error adding a penalizing factor to the popular MAD criteria is adopted since this allows increasing the search range from 8 to 32, regularizing the motion vector field by favoring 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, the motion vector that intersects closer to its center is selected.

IV. Creating Inter-View Side Information

To exploit the inter-view correlation, an estimation of the WZ frame has to be made based on the past pairs of decoded frames, using both first and second views. This disparity-based frame estimation (DBFE) can be made through:

1. Estimation of the disparity field for a past pair of frames.
2. Extrapolation of the inter-view side information for the second view under decoding by applying the estimated disparity field to the first view decoded frame which generates an estimate of the second view.

While there are many disparity estimation algorithms in the literature, the one adopted here corresponds to the motion estimation presented in subsection III.B. While this

may not be the best disparity estimation solution, it simplifies the decoder by reusing the same algorithm as for temporal interpolation. While disparity estimation is not the core of this paper, the proposed architecture (and the associated software implementation) allow to easily plug in more sophisticated disparity estimation methods, notably depending on the cameras position, and the scene geometry [9].

V. Creating Fused Side Information

Since the solutions previously presented to create the stereo side information only exploit a single correlation dimension (temporal or inter-view), it is natural to expect that a side information creation solution exploiting both correlation dimensions may lead to more efficient WZ coding. For this to happen, it is necessary to find an efficient way to fuse the temporal and inter-view correlations in order the decoder may at least take benefit of the most powerful of them for each decoded frame.

A. Proposing a Mask-Based Fusion Approach for WZ Stereo Video Coding

In this paper, a fusion-based approach is proposed based on the following ideas:

1. For each WZ frame under decoding, the (fused) side information is created based on the temporal and inter-view side informations; this side information which is expected to be better than each individual side information should provide better RD performance than temporal or inter-view alone.
2. The process to fuse the two individual side informations is based on a binary fusion mask created after decoding the most recent frame; pixel by pixel, this mask is set to '1' if the temporal side information is the most similar to the decoded frame and set to '0' if inter-view side information is the most similar.

The proposed architecture for the mask-based fusion solution for WZ stereo video coding is presented in Figure 3. In this architecture, temporal side information is created using the motion-based frame interpolation solution described in Section III while inter-view side information is created using the disparity-based frame extrapolation solution described in Section IV.

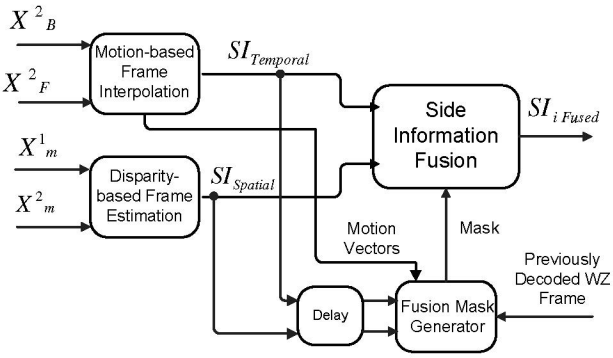


Figure 3 – Stereo side information generation architecture using a mask-based fusion approach.

B. Filling the Fusion Mask

The binary decision mask used for each WZ frame being decoded indicates which is the best side information to use for each pixel: the temporal or inter-view side information. Both SI frames were already created for the current time instant t using the algorithms presented in sections III and IV. To create this mask, three solutions are proposed:

1) Ideal Mask

The ideal mask solution is here presented to be taken as reference since it gives an upper bound for the RD performance to be obtained by fusing the available temporal and inter-view side informations (which depend on the 'quality' of the motion estimation and disparity estimation solutions adopted). The ideal mask is computed using the original frame instead of the decoded frame, and setting the mask to '1' if the temporal side information is closer to the original frame and vice-versa for the inter-view side information.

2) Delayed Mask

For this solution, the fusion mask is created after the decoding of the most recent WZ frame before the one under decoding. The delayed mask is computed using the decoded and both side information frames for the most recent WZ decoded frame, setting the mask to '1' if the temporal side information is closer to the decoded frame and vice-versa for the inter-view side information. This solution has the drawback that the fusion mask is delayed in time since the decision was performed in the past ,i.e. before the current decoding time t, e.g. two frame periods (t-2) for GOP=2.

3) Motion Compensated Mask

This last solution tries to overcome the limitations of the delayed mask solution by motion compensating the fusion mask from the last decoded WZ frame using the motion field estimated for the current WZ frame, using the process defined in Section III and assuming linear motion. In this way, it is expected the fusion mask to be more accurate and closer to the ideal mask.

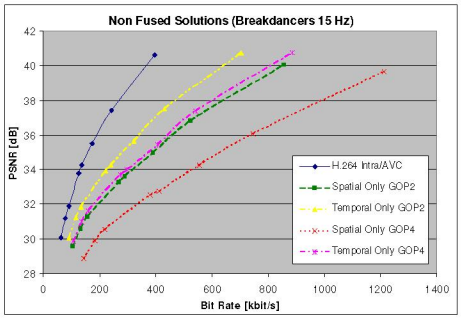
VI. Experimental Results

This section evaluates the rate-distortion (RD) performance for all the WZ stereo coding architectures previously defined under adequate conditions. The tests carried out used two sequences, notably Breakdancers and Uli. All frames were used for all sequences, which means 99 frames for Breakdancers (15 Hz), and 249 frames for Uli (25 Hz) at 256×192 luminance spatial resolution. Both sequences have 8 different views available; for the stereo tests here reported, the first and second views have been used.

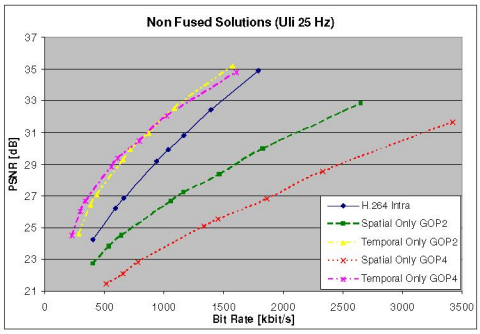
The TDWZ monoview codec was configured to use a fixed GOP size, and quantization matrices as defined in [5]. The key frames were encoded with H.264/AVC Intra (Main Profile). As usual for WZ coding, only luminance data has been coded. Bitrate includes the luminance rate for WZ frames and key frames for the second view to be coded since the first view is always the same. This allows evaluating the RD performance of the 2nd view coding assuming the first has been coded.

This section reports and analysis the RD performance for the following cases:

1) Non fused Wyner-Ziv coding: Figure 4 shows the results for GOP 2 and GOP 4 using only temporal or inter-view side (spatial) information. It is clear that temporal side information leads to better RD performance since the MBFI algorithm is more mature than the DBFE and there is typically more temporal than inter-view dependency. For inter-view side information, GOP 4 results are worst than GOP 2 results because for this sequence key frames (H.264/AVC Intra coded) are still more efficient than inter-view WZ coding.



a)



b)

Figure 4 – RD performance for non-fused side information: a) Breakdancers; b) Uli.

2) Fused Wyner-Ziv coding: Figure 5 shows the results for the various fused coding solutions. The first important conclusion is that, for all tested conditions, the ideal mask RD performance is always better than the temporal only WZ coding which shows that there is dependency to exploit and to take benefit from the first view. The fact that the 'real mask' is either delayed or based on basic motion extrapolation explains its current poorer performance but it is clear that there are large margins for improvement. This margin increases with the GOP size since the larger is the GOP size the poorer is the temporal side information (the key frames are less correlated) and thus it is more important to exploit the inter-view dependency.

VII. FINAL REMARKS

This paper proposes a new approach to Wyner-Ziv stereo video coding based on temporal and inter-view side information fusion. The results clearly show the potential of the side information fusion approach. The current results are limited by two main factors: i) the less mature disparity estimation approach; and ii) the delayed or too simply motion compensated fusion mask. The next research will be concentrated on these two topics.

REFERENCES

[1] J. Slepian, and J. Wolf, "Noiseless Coding of Correlated Information Sources", *IEEE Trans. on Inform. Theory*, Vol. 19, No. 4, pp. 471-480, July 1973.

[2] A. Wyner, and J. Ziv, "The Rate-Distortion Function for Source Coding with Side Information at the Decoder", *IEEE Trans. on Inform. Theory*, Vol. 22, No. 1, pp. 1-10, January 1976.

[3] M. Flierl, and B. Girod, "Coding of Multi-View Image Sequences with Video Sensors", *IEEE ICIP*, Atlanta, GA, USA, October 2006.

[4] M. Ouaret, F. Dufaux, and T. Ebrahimi, "Fusion-Based Multiview Distributed Video Coding", *ACM Intl. Workshop on Video Surv. and Sensor Networks*, Santa Barbara, CA, USA, October 2006.

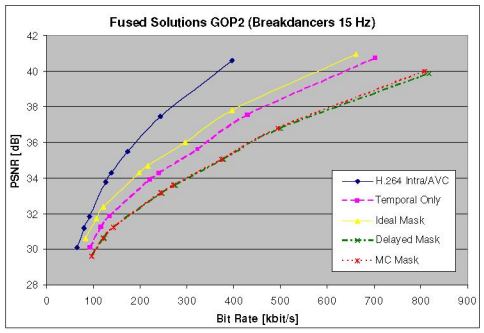
[5] C. Brites, J. Ascenso, and F. Pereira, "Improving Transform Domain Wyner-Ziv Video Coding Performance", *IEEE ICASSP*, Toulouse, France, May 2006.

[6] 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 and Image Processing, Multim. Com.and Services*, Smolenice, Slovak Republic, June 2005.

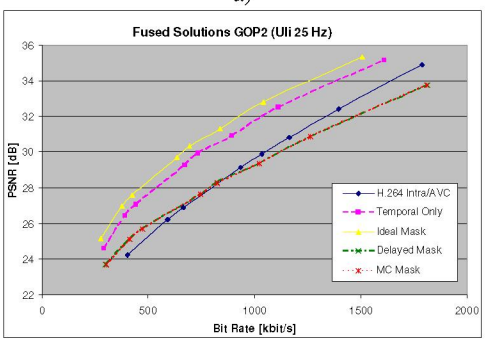
[7] A. Aaron, S. Rane, E. Setton, and B. Girod, "Transform-Domain Wyner-Ziv Codec for Video", *VCIP*, San Jose, USA, January 2004.

[8] J. Ascenso, C. Brites, and F. Pereira, "Content Adaptive Wyner-Ziv Video Coding Driven by Motion Activity", *IEEE ICIP*, Atlanta, USA, October 2006.

[9] X. Huang and E. Dubois, "Disparity Estimation for the Intermediate View Interpolation of Stereoscopic Images", IEEE ICASSP, Philadelphia, PA, USA, March 2005.



a)



b)

Figure 5 – RD performance for fused side information: a) Breakdancers GOP 2; b) Uli GOP 2.