Distributed Video Coding for Wireless Multi-Camera Networks

Nantheera Anantrasirichai, Dimitris Agrafiotis, Dave Bull

University of Bristol, Woodland Road, Bristol, BS8 1UB, UK

Keywords: Multiview, distributed video coding, surveillance.

Abstract

In this paper, we present a novel distributed multiview video compression framework for surveillance purposes that employs a concealment based approach for the generation of the side information through the use of hybrid Key/Wyner-Ziv frames. Two multi-view coding structures are proposed in order to better exploit the inter-camera correlation present at the decoder. We additionally introduce a number of enhancements including use of spatio-temporal concealment for generating the side information on a MB basis, mode selection for switching between the two concealment approaches and for deciding how the correlation noise is estimated, local (MB wise) correlation noise estimation and modified B frame quantisation. The results presented indicate considerable improvement (up to 23%) compared to corresponding frame based schemes and improvements of up to 1.4 dB relative to an equivalent intra-view codec.

1 Introduction

Low cost - low power video encoders can play an important role in multi-camera wireless video surveillance systems, where limited availability of, potentially irreplaceable, power sources can render the complexity/compression trade off more relevant than pure rate-distortion performance. The latter is normally the goal of the power hungry motion estimation based hybrid codecs, including the latest H.264/MPEG4-10 video coding standard. Furthermore, in such wireless surveillance systems, communication among cameras is not always feasible or desirable, as it adds significant strain to the network's resources and can create dependencies among cameras making the whole system even more vulnerable to the already present channel errors as well as to malfunctions etc.

Distributed video coding (DVC) has recently received considerable interest as an approach to video coding that offers an alternative solution to the complexity balance issue between the encoder and the decoder. DVC allows shifting the complexity from the encoder to the decoder making it a particularly attractive approach for low power systems with multiple remotely located encoders, such as multi-camera wireless video surveillance and multimedia sensor networks. DVC stems from information theory results developed in [1] by Slepian and Wolf, and extended in [2] by Wyner and Ziv, on source coding with side information at the decoder. They effectively prove that it is possible when two statistically dependent signals X and Y are considered, to compress X when Y (known as the side information) is available only at the decoder at a rate similar to the case where Y was available at the encoder. In the multi-view scenario discussed herein, where several cameras capture the same scene from different angles, DVC additionally offers the possibility of exploiting spatial/view correlations among cameras without the need for them to communicate, since the side information (i.e. the prediction) is formed at the decoder side.

The most common approach to DVC is that where the frames of a single source video are split into two categories, key frames and WZ frames [4][5][6]. Key frames are intra coded with a conventional encoder. WZ frame coding, which may involve transformation, uses quantisation followed by channel coding applied in a bitplane by bitplane fashion, with the parity bits only being transmitted to the decoder. At the decoder, the key frames are used for creating an estimate of the WZ frames (side information). This side information (SI) is seen as the systematic part of the channel encoder's output as received at the decoder, i.e., the SI is seen as a noisy version of the original coded WZ frame. The received parity bits are employed to correct the errors presenting in this noisy version of the WZ data. An excellent list of relevant papers can be found in [7].

In previous work of ours [8][9] we proposed the use of Hybrid Key/Wyner-Ziv frames (KWZ) for better generation of the SI and more accurate estimation of the correlation noise. According to this approach the process of generating the SI is treated as a concealment task. By employing a macroblock pattern similar to the one specified by the dispersed flexible macroblock ordering (FMO) of H.264, we group the macroblocks of each frame into intra coded (key) and Wyner-Ziv groups. Temporal concealment is then used at the decoder for "concealing" (predicting) the missing WZ macroblocks using the information available from the already received 4-neighboring key MBs. The same key MBs are also used for estimating the correlation noise through motion estimation for the whole WZ group. This approach was tested and found to offer performance advantages relative to the more common approach of splitting the sequence in key and WZ frames, wherein SI generation involves concealing / recovering a whole missing frame; a process normally associated with poor quality results. Moreover in frame based approaches the estimation of the correlation noise relies on predicted and not actual received data.

In this paper, we extend our previous DVC scheme to the case of multi-view sequences, where the correlation among adjacent cameras is also exploited in order to construct the SI more accurately. Occluded and revealed areas which can not be accurately predicted from temporally adjacent frames of

Printed and published by the IET,
Michael Faraday House, Six Hills Way, Stevenage, Herts SG1 2AY

the same view can potentially be visible/present and hence better predicted from neighbouring views. As a result the reconstructed frames can be of better quality using what is effectively intra coding at the encoder side. We introduce a number of enhancements at the decoder for improving the performance of our codec: i) use of spatio-temporal concealment for generating the SI on a MB basis; ii) local (MB wise) correlation noise estimation; iii) modified B frame quantisation. The results indicate considerable performance improvement relative to other frame based schemes, including our own multi-view frame based codec presented in [10].

The rest of this paper is organized as follows: Section 2 reviews existing multi-view DVC techniques. The proposed scheme is described in Section 3. Section 4 presents results and comparisons and Section 5 concludes this paper.

2 Previous work

In multi-view systems, several cameras capture the same scene from different angles. Sequences therefore present intra-view (temporal and spatial) correlations as well as inter-view redundancies. Early work on multi-view DVC [11],[12] focused on systems with three cameras, with two of the cameras using pure (H.264) intra coding and one camera using DVC techniques. Fusion methods were introduced in [12] to merge the intra- and inter-camera estimations made at the decoder. In [13], DVC coding for a large camera array was proposed with some cameras using pure intra coding, and the rest exploiting these intra-coded views through geometry based rendering to generate their SI. An alternative coding structure was proposed in [14], where all views are DVC coded with the Key and WZ frames alternating both in time and space. This allows the WZ frames to be reconstructed using temporally (intra-view) and spatially (inter-view) adjacent frames. However, on the assumption of a linear motion trajectory and linear view transformation, the performance of these multiview DVC schemes is still far from the hybrid prediction schemes.

3 Proposed Scheme

The first step in the proposed framework involves splitting of the current frame into Key and WZ groups of macroblocks, in a similar fashion to the dispersed FMO specified in H.264 (see Figure 2). Each Key group is encoded with H.264 in Intra mode. If transform-domain coding is applied, the WZ MBs are first transformed, using the same DCT like transform that is employed in H.264. The WZ MBs (pixels or coefficient values) are then quantised and bit-planes of the quantised symbols are extracted. These are then fed to the turbo encoder and parity bits are produced which are stored in a buffer. At the decoder, the SI is created using temporal and spatial error concealment (TEC/SEC) methods. The turbo decoder uses the parity bits and the (possibly transformed) SI to form the decoded bit-planes. If transform-domain coding is enabled, the reconstructed WZ group is then inverse-transformed. Finally, the decoded WZ group and the decoded Key group are merged and a de-blocking filter is applied to remove blocking artefacts occurring between WZ and Key MBs. The proposed scheme is described in more detail below.

3.1 GOP Structure

The structure of the intra-view group-of-pictures (GOP) employed in our scheme is similar to the one typically used in hybrid coding schemes with IBP frames. A fully intra coded Key frame is placed at the first frame of each GOP with the subsequently decoded P frames employing the two most recently decoded frames (Key and KWZ) as references for generating the SI. The two MB groups alternate from one P frame to another P frame so as to avoid creating potentially annoying regions of different subjective quality. The MB groups of the B frames alternate relative to the previous (in display order) reference frame.

The multi-view coding structure influences the performance of the system as it affects the side information generation. The available reference frames for decoding the current frame of a specific view/camera depend on how intra and WZ/ KWZ frames alternate in time and space. We propose two coding structures as shown in Figure 1, where the numbers in brackets indicate the coding order. For structure 1, the assumption that the two furthest views can better cover the whole scene is utilised so that the WZ blocks in the middle views can have at least one visible reference. In structure 2 on the other hand, the middle views are coded before the views at the edges in order to better exploit the high redundancy normally present between closely located adjacent views. Structure 2 allows slightly faster decoding as the number of reference frames employed is smaller compared to structure 1.

3.2 B-Frame Quantisation

The SI generation for the WZ macroblocks of the B frames relies on bidirectional error concealment that utilises multiple reference frames for predicting the "missing" MBs. As a result the SI for the WZ MBs of these frames is more reliable and closer to the transmitted data resulting in a lower bitrate for these frames. In order to further reduce this bitrate we increase the quantisation step size of these WZ MBs relative to the step size used for the WZ MBs of the P frames. As in the scalable extension of H.264 (SVC) [15] the associated PSNR fluctuation within the GOP should not appear subjectively annoying as long as the number of quantisation levels is chosen appropriately. The relationship between the number of quantisation levels Q_B^n and Q_P^n for B and P frames respectively, for n bit-planes can be written as:

$$Q_B^n = \left| a \cdot Q_P^n \right| \tag{1}$$

Based on experimental results, we have selected a so that $Q_B^n = 1/2(Q_P^n + Q_P^{n-1})$ i.e. $a = 0.75$.

3.3 Side Information Generation

The generation of the SI is equivalent to an error concealment process for missing macroblocks (WZ MBs) in the presence of all their 4-neighbours. We employ temporal and spatial error concealment methods (TEC/SEC) the application of which is controlled by a mode selection algorithm [16].

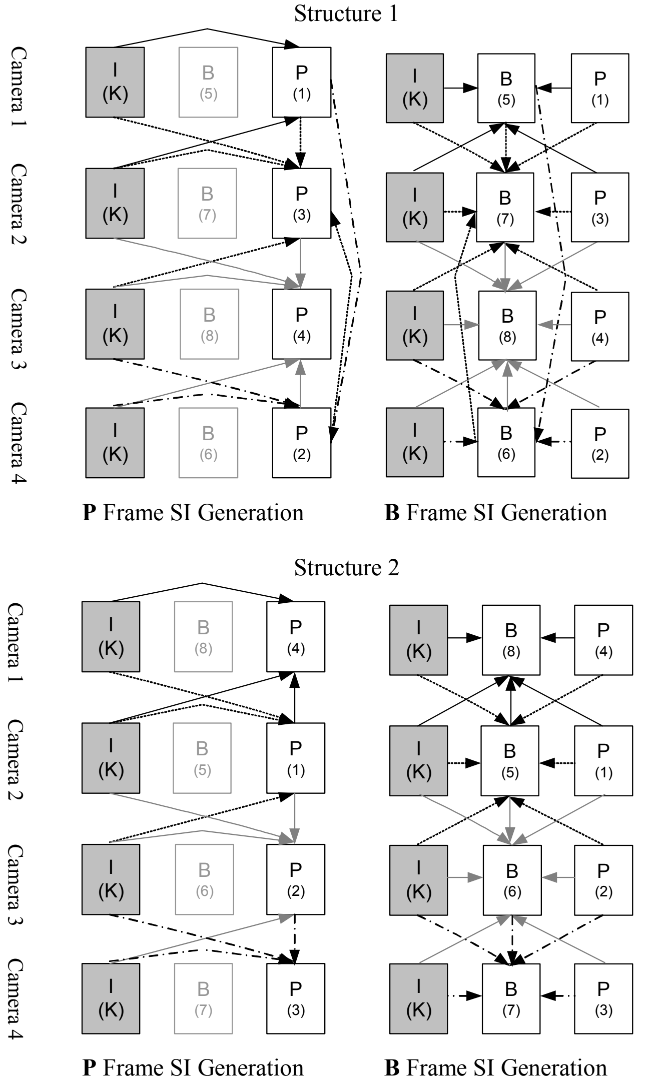

Figure 1: Proposed multi-view coding structures showing coding order and reference frames available for P and B frame SI generation

The employed TEC method uses the external boundary matching error (EBME) of a WZ MB -defined as the sum of absolute differences between the multiple pixel boundary of MBs adjacent to the missing one (WZ) in the current frame and the same boundary of MBs adjacent to the replacement MB in the reference frame (Figure 2) - in order to test possible motion vectors (MV_{WZn}) from a list that includes those of spatially and temporally adjacent MBs as well as the zero MV. Spatially adjacent MVs are generated through forward (and backward for the case of B frames) motion estimation for all 8x8 blocks of the Key MB group (MV_{Kn}) prior to the concealment process. In the multi-view case this motion estimation takes place also in reference frames coming from other views. The search range of the motion estimation is adjusted according to the distance between the current and the reference frame. The initial estimation of a WZ MV is further refined through overlapped block motion compensation. In the case of B frames the replacement (SI) macroblock can also result from averaging a forward and backward replacement MB depending on the EBME.

The SEC module uses bordering Key pixels to conceal the missing WZ pixels of each WZ MB through bilinear interpolation or directional interpolation along detected edges. The type of interpolation used depends on the outcome of a decision algorithm that uses the directional entropy of neighbouring edges for choosing between the two interpolation approaches.

The mode selection algorithm examines the suitability of the TEC method for concealing each WZ MB, by evaluating the levels of motion compensated activity and spatial activity in the neighbourhood of that MB and switching to spatial concealment accordingly. Motion compensated temporal activity is measured as the mean squared error between the key MBs surrounding the missing one in the current frame and those surrounding the replacement MB in the reference frame. Spatial activity is measured as the variance of the surrounding key MBs in the current frame. More formally:

$$SA = \mathrm{E}\,[\,(x-\mu)^2\,] \text{ and } TA = \mathrm{E}\,[\,(x-x^*)^2\,] \qquad (2)$$

where x are the pixels in the neighbourhood of the missing MB and x^* are the pixels in the neighbourhood of the replacement MB in the reference frame. SEC is employed if the spatial activity is smaller than the temporal activity and the latter is above a specific threshold (3 in this work). Otherwise TEC is used.

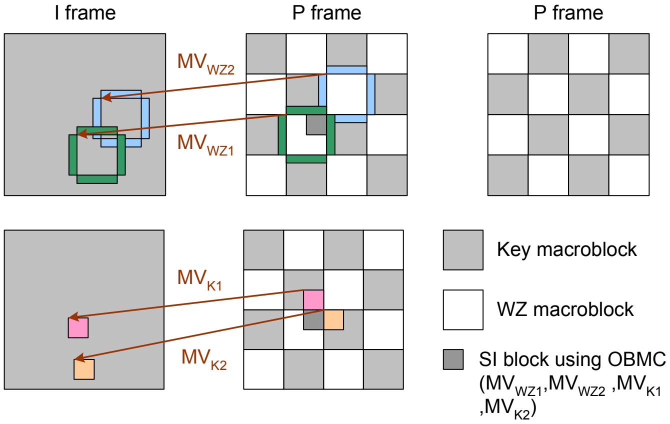

Figure 2: Example of the TEC process, as applied to a P frame.

3.4 Correlation Noise Estimation

We estimate the correlation noise on a macroblock basis using the 4-neighbouring Key macroblocks of each WZ MB. We model the noise as a Laplacian distribution with a specific variance that changes from MB to MB. If TEC is selected, we employ the difference between the 4-neighboring key MBs and the corresponding motion compensated MBs in previously decoded frames that were found to provide the best match during the motion estimation process, in order to estimate the Laplacian distribution parameter-α. In other words, after having performed motion estimation for the key MBs of the current frame as described in section 2.3 we take the difference between each key pixel and its best match (as indicated by ME) in one of the reference frames. The resulting distribution should follow closely that of the difference between the SI MB and the transmitted WZ MB as it employs actual received pixels in the vicinity of the processed MB, as opposed to frame based interpolated values. If SEC is applied for creating the SI MB then α is calculated using the variance of the difference between the 4-neighbouring Key MBs and this SI MB.

4 Results and Discussion

The performance of the proposed codec was evaluated using the well known test multi-view sequence *"Breakdancers"* (Microsoft) at CIF resolution. The proposed codec was compared to the multiview DCV scheme of [10] without the use of depth maps (simple block matching between views). PD-DVC and TD-DVC stand for pixel and transform domain DVC respectively. Str1 and Str2 indicate the use of structure 1 and structure 2. The reported bitrate and PSNR values are average values for all views and include both key and WZ data. Mono view results (referred to as "mono" in Figure 3) make use of intra-view information only.

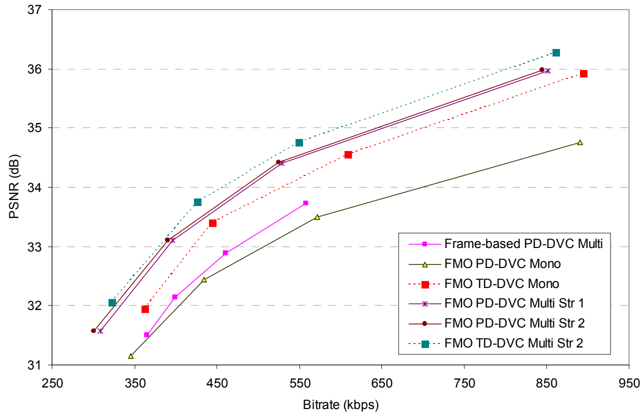

Figure 3: Performance evaluation of proposed system.

The results indicate that the proposed scheme outperforms existing frame based schemes in the pixel domain. Adding a transform leads to further improvements (up to 23% bit rate reduction). The use of inter-view data improves the performance of the system significantly as shown by the performance difference between FMO PD-DVC Mono and Multi (~1.4 dB). The proposed multiview structures show insignificant different results but we suggest that more multiview test sequences should be used as the wide- and narrow-baseline geometries could affect the performance of the codec.

5 Conclusions

A novel distributed multiview video coding framework based on block-based error concealment is proposed in this paper. Hybrid Key/WZ frames are employed via an FMO type interleaving of macroblocks. Our approach allows better SI generation and more accurate correlation noise estimation, both performed at the MB level. Intra- and inter- view data is used for generating the side information, through the introduction of two coding structures. Results with the proposed method show considerable improvement compared to corresponding frame based schemes (up to 23% reduction in bitrate).

Acknowledgement

This work was done as part of the EPSRC funded grant EP/E065538.

References

[1] D. Slepian and J. K. Wolf, "Noiseless Coding of Correlated Information Sources," IEEE Transactions on Information Theory, vol. 19, no. 4, July 1973.

[2] A. D. Wyner and J. Ziv, "The Rate Distortion Function for Source Coding with Side Information at the Decoder," IEEE Transactions on Information Theory, vol. 22, no. 1, January 1976.

[3] B. Girod, A. Aaron, S. Rane, D. Rebello-Mondero, "Distributed Video Coding," Proc. of the IEEE, vol. 93, no. 1, Jan. 2005.

[4] A. Aaron, R. Zang, and B. Girod, "Wyner-Ziv Coding of Motion Video," in ASILOMAR Conf. on Signals and Systems, Nov. 2002.

[5] J. Ascenso, C. Brites, and F. Pereira, "Improving Frame Interpolation with Spatial Motion Smoothing for Pixel Domain Distributed Video Coding," in EURASIP, Video/Image Processing and Multimedia Communications, June 2005.

[6] C. Brites, J. Ascenso, F. Pereira, "Improving Transform Domain Wyner-Ziv Video Coding Performance", IEEE ICASSP, Toulouse, France, May 14-19, 2006

[7] The DISCOVER project, http://www.discoverdvc.org/

[8] D. Agrafiotis, P. Ferré, D. R. Bull, "Hybrid key/Wyner-Ziv frames with flexible macroblock ordering for improved low delay distributed video coding", VCIP 2007, 28 January–1 February 2007, San Jose, California.

[9] N. Anantrasirichai, D. Agrafiotis, D. R. Bull, "A Concealment Based Approach To Distributed Video Coding", submitted to ICIP 2008

[10] P. Ferre, D. Agrafiotis, D. R. Bull, "Fusion Methods for Side Information Generation in Multi-View Distributed Video Coding Systems", ICIP 2007, Sept. 16-19 2007.

[11] X. Artigas, E.Angeli, L. Torres, "Side Information Generation for Multiview Distributed Video Coding Using a Fusion Approach," 7th *Nordic Signal Processing Symposium*, June 2006.

[12] M. Ouaret, F. Dufaux, T. Ebrahimi, "Fusion-based Multiview Distributed Video Coding," in *ACM Int. Works. Video Surveillance and Sensor Networks*, 2006.

[13] X.Zhu, A. Aaron, B.Girod, "Distributed Compression for Large Camera Arrays," in *Proc. IEEE Workshop on Statistical Signal Processing,* 2003, pp. 30-33.

[14] X. Guo, Y. Lu, F. Wu, W. Gao, "Distributed multi-view video coding," *SPIE Visual Communications and Image Processing*, VCIP 2006, San Jose, CA, USA, Jan. 2006

[15] H. Schwarz, D. Marpe, T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," IEEE Trans. on Cir. and Syst. for Video Technology, vol.17, no.9, pp.1103-1120, Sept. 2007

[16] D. Agrafiotis, D. R. Bull, N. Canagarajah, "Enhanced error concealment with mode selection", IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 8, pp. 960-973, August 2006.