ENHANCED SPATIALLY INTERLEAVED TECHNIQUES FOR MULTI-VIEW DVC

Nantheera Anantrasirichai, Dimitris Agrafiotis

Department of Electrical & Electronic Engineering,
University of Bristol, Bristol, BS8 1UB, UK

ABSTRACT
This paper presents a novel distributed multi-view video coding framework with independent camera encoding and centralised decoding. Spatio-temporal-view concealment methods are employed for generating the side information (SI) through the use of hybrid Key/Wyner-Ziv frames. We apply a diversity technique for fusing multiple SI thereby achieving more reliable results. We additionally introduce two enhancements for further improving the rate distortion performance of the system, namely static area detection and inter-view bitplane projection. Results show a significant improvement in performance relative to H.264 intra coding of up to 25% reduction in bitrate (or equivalently 2.5 dB in PSNR).

Index Terms— DVC, Multi-view Video coding

1. INTRODUCTION
Distributed video coding (DVC) [1,2] 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. 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 of communication, since the prediction is performed at the decoder side.

A common DVC scenario involves splitting of the video frames into two categories, Key frames and Wyner-Ziv (WZ) frames [3]. Key frames are conventionally intra coded while, WZ frame coding, which may involve transformation, uses quantisation followed by channel coding applied in a bitplane-by-bitplane fashion, with only the parity bits being transmitted to the decoder. At the decoder, the decoded Key frames are used to create an estimate of the WZ frames. The received parity bits are then employed to correct the errors occurring in this noisy version of the WZ data.

In this paper, we employ our previously proposed hybrid Key/WZ frame approach whereby spatial interleaving (SPI) of Key and WZ blocks takes place using a chessboard pattern similar to the flexible macroblock ordering (FMO) dispersed pattern of H.264 [4,5]. The results indicate performance improvements relative to frame based schemes [6], especially for sequences with fast or non-linear motion.

In order to avoid any performance loss relative to full frame Key coding, we additionally employ temporal interleaving whereby two Key groups of two consecutive frames are interleaved prior to intra coding. For WZ groups, we use a Gray code which has been shown to improve system performance [7].

At the central decoder, encoded data streams from all cameras are decoded together using block-based concealment to “conceal” (predict) the missing WZ blocks using information available from the previously received 4-neighbouring key blocks. Generally one of the difficulties of multi-view DVC is defining a SI fusion method, as several SI can be generated using temporally (intra-view) and spatially (inter-view) adjacent frames; however, no original frame exists to enable comparison. In this paper, we introduce a simple and efficient fusion method inspired by a diversity technique used in wireless communications. In the DVC scenario this process translates into a multi-hypothesis SI framework. Multiple SI data generate a more precise log-likelihood ratio thereby reducing the number of bits requested by the turbo decoder for correcting the SI [8].

We, moreover, propose two methods to enhance the performance of our multi-view (MV) codec: static area mark and bitplane projection. Both techniques offer significant reductions in bitrate. The first method codes only the data at the moving areas. The second method exploits the associated bitplanes among views. The static area mark technique requires a slight modification of the encoder for detecting the static areas, whereas the bitplane projection technique can make use of traditional encoders. Decoding using the bitplane projection method is more complex due to the depth estimation process involved. Although the overall complexity of the system is increased, the results show considerable improvement over the traditional codec.

The rest of paper is organised as follows: Section 2 describes the proposed codec and the enhancement techniques. Results and discussion are presented in Section 3. The conclusions are presented in Section 4.

2. PROPOSED MULTI-VIEW SPI-DVC
The proposed MV SPI-DVC uses an IB coding structure.
The first frame of each GOP is coded using Intra-frame coding (I-frame), while other frames of the GOP are coded using a combination of Key and WZ coding, as well as interpolation for generating the SI (B-frame). To maximise data-prediction efficiency and minimise error accumulation, the distance between the B-frame and the I-frames is globally minimised by shifting the positions of the GOPs in the even views. That is, intra-frame coding in these views is applied to the first frame and frames \( (\frac{GOP+1}{2}) + n \cdot GOP \), \( n \in \{0,1,\ldots\} \). The B-frame with the shortest distance to I-frames is decoded first. Sample GOPs of 2 and 3 are illustrated in Fig. 1. The proposed codec is explained below.

2.1 SPI-Encoder

As communication among cameras is unavailable, each camera view is independently encoded. For general details regarding the encoder we refer readers to [4]. Below we only describe the key proposed features.

2.1.1 Spatio-Temporal Interleaving

The first step involves splitting of the current input frame into Key and WZ groups in a similar fashion to the dispersed type of FMO specified in H.264. The interleaving block size can range from 16x16 to 4x4 pixels. If the GOP length is more than 2, the Key and WZ groups alternate relative to the previous frame so as to avoid creating potentially annoying regions of different subjective quality. The Key groups of two successive frames are combined to avoid any significant performance loss, especially in the case where the complex spatial prediction modes of H.264 are used, relative to full frame intra KEY coding. The same procedure is applied to the WZ groups in order for the frame length of the input to the turbo encoder to be adequate for good performance.

2.1.2 A Gray Code

After transformation and quantisation, the quantised symbols are converted into binary data. Subsequently, the binary codes are converted to Gray codes by XORing the binary values with their logical-shift-right values. At the decoder, the decoded Gray codes are converted back to binary values in a similar fashion. The Gray code conversion clearly adds a small complexity to the system, but improves the codec’s performance as it has a Hamming distance of one which provides higher error resilience than the natural binary code, and offers a higher accuracy of bit probability estimation.

The WZ data are fed to a Turbo coder in a bitplane-by-bitplane fashion, and then the parity bits are transmitted depending upon the request from the decoder.

2.2 Multi-view SPI-Decoder

The decoder works as a central base station. The received data from all cameras are jointly decoded. Again, due to space limitations we refer the readers to a previously published paper of ours for more details regarding the decoder [9]. Key proposed features in this paper are described below.

2.2.1 Side Information Generation

The generation of the SI is equivalent to an error concealment process for missing blocks (WZ blocks) in the presence of their 4-neighbours (Key blocks). We employ temporal (or disparity when inter-view references are used) and spatial error concealment methods (TEC/SEC) controlled by a mode selection algorithm as proposed in [10] for generating the SI. For TEC, each missing WZ block is divided into four sub-blocks. Each sub-block is concealed using two motion vectors coming from neighbouring Key blocks and two motion vectors generated using an external boundary matching error process, fused via a cosine weighted overlapping step. The SEC module uses bordering Key pixels to conceal the WZ blocks through bilinear or directional interpolation depending on the directional entropy of neighbouring edges. The mode selection algorithm switches between TEC and SEC based on the levels of motion compensated activity and spatial activity in the neighbourhood of the processed block. If forward and backward reference frames or left and right references are available, bi-directional motion/disparity estimation is employed so that the replacement (SI) block can additionally result from averaging a forward and backward replacement block.

2.2.2 Multi-hypothesis Decoding

The multi-hypothesis approach proposed in [5] has been applied to the MV DVC codec by exploiting multiple SI generated using motion and disparity estimation. In this paper we use three SI data; \( SI_2 = \) intra-view SI; \( SI_3 = \) inter-view SI; \( SI_1 = \) best estimation between \( SI_2 \) and \( SI_3 \). As
multiple SI data are available, multiple bit probabilities can lead to a more precise log-likelihood ratio (LLR). Moreover, having multiple SI data allows usage of the correct ones to compensate for errors appearing in the others, so that the quality of the final reconstructed image is significantly improved.

2.3 Enhancement techniques for MV-DVC

In this paper we propose two enhancement techniques applied to the multi-view scenario; static area mark and bitplane projection.

2.3.1 Static Area Mark

Most multi-camera systems are installed with fixed camera positions. As a result, a video sequence captured in each camera displays the same background and moving objects. If cameras have the ability to detect the static background, this redundant data need not be transmitted. The proposed encoder simply extracts the static areas with a subtraction technique. For each macroblock of a B frame consisting of two Key blocks and two WZ blocks, a sum of absolute difference (SAD) between two successive frames is compared with a predefined threshold. Macroblocks with SAD above the threshold are defined as moving areas and are consequently encoded.

Fig. 2 demonstrates the proposed static area mark scheme. A mark map is created with ‘1’ for moving (dynamic) macroblocks and ‘0’ for static macroblocks. The dynamic macroblocks of the first frame of two consecutive B frames are split into Key and WZ blocks. Then, the dynamic Key blocks are overlaid on the reference I frame to create a new reference frame for the second B (B2) frame. This new reference should provide a better prediction for B2 as it has been updated with the more recent dynamic key blocks. After detecting the static blocks of the second B frame (B2), the two sets of dynamic WZ data are combined and coded together via WZ coding. The two mark maps generated can be transmitted with or without prior encoding (e.g. JBIG). Two sets of Key data are also combined to create a new Key frame. The empty blocks, which correspond to the static blocks not being sent, are filled with the mean value of the dynamic blocks. This Key frame is then encoded with H.264 Intra coding. At the decoder, the blocks marked as dynamic (marked with 1) are reconstructed as explained in Sec 2, while the rest of the blocks are replaced by the co-located blocks in the reference frame.

2.3.2 Bitplane Projection

The aim of this technique is to reduce the bits transmitted. Some cameras are selected to transmit only the Key data: Key frames and Key slice groups. At the decoder the B-frames of these cameras are reconstructed using depth information, Key data and reconstructed frames from other views. Note that as at least two views are conventionally coded, the decoder has enough information to estimate the depth of the scene. Note however, that in this paper we used the depth maps provided with the test multi-view video sequences. These depth maps have been generated using the original (lossless) video. In a real scenario, the depth would be estimated from the reconstructed views which can suffer from artefacts and high-frequency loss. The effect of that will be studied in the future.

The SI of the WZ data is generated using common techniques, but is corrected using decoded bitplanes projected from neighbouring views instead of receiving parity bits from the encoders. The reliability of the projected bitplanes is checked, since the estimated depth possibly contains errors. We propose two approaches of checking reliability; an intensity/depth consistency or a first BP agreement.

For the intensity/depth consistency, two thresholds are set (here 15 and 4 for intensity and depth consistencies respectively). If the differences of the intensity and depth between the projected view and the Key data available at the neighbouring blocks of a particular WZ block in the current view are within the thresholds, the projected bitplanes are used to reconstruct the WZ block.

A more reliable approach can be applied if the first WZ bitplane is decoded. The decoder requests the parity bits of this bitplane, and then uses the reconstructed bitplane to
check whether the projected first bitplane is identical. If so, the rest of the projected bitplanes are employed. Note that bitplanes of the adjacent views can be used in the current view as it is captured from the same scene.

3. RESULTS AND DISCUSSION

The performance of the proposed codec was evaluated using four views of two test multi-view sequences; Breakdancers and Ballet at CIF resolution. The results were compared to H.264 Intra coding and H.264 predictive coding with zero search range (IBI mv-zero coding). Fig. 3 top and bottom show rate-distortion plots of Breakdancers and Ballet respectively. It is clear that exploiting the correlation in both temporal and inter-view directions improves the rate distortion performance of the codec, with a reduction in bitrate of up to 18% and 35% for Breakdancers and Ballet respectively. The bitplane projection technique offers limited improvement at low bitrates. If the parity bits for the first bitplane (BP proj+1BP) are additionally requested, then the performance loss at higher rates is mitigated. The plots show interesting results of which the performance is close to H.264 IBI mv-zero coding, particularly at low bitrates. Clearly the proposed schemes can significantly reduce transmission data with small deterioration in reconstructed video quality compared to the traditional SPI DVC.

4. CONCLUSIONS

This paper presented a novel distributed multi-view video coding framework with Hybrid Key/WZ frames employed via spatio-temporal interleaving of blocks. Each view is encoded completely independently from the others, while the centralised decoder reconstructs multi-view video sequences based on block-based error concealment. We proposed two enhancement techniques for improving the performance at the cost of a slightly increased system complexity. The first technique detects redundant data between frames and the second one employs bitplanes from neighbouring views. Results show significant improvement over H.264 Intra coding and comparable performance with H.264 predictive coding with zero motion.

5. REFERENCES