FUSION METHODS FOR SIDE INFORMATION GENERATION IN MULTI-VIEW DISTRIBUTED VIDEO CODING SYSTEMS

Pierre Ferré, Dimitris Agrafiotis, David Bull

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

ABSTRACT

The generation of the side information is an important part in the design of a distributed video coding (DVC) system as it directly relates to the system's rate distortion performance. In multi-view systems spatial/view inter-camera correlations can be exploited alongside temporal/motion intra camera ones for generating the side information as accurately as possible. In this paper, algorithms for fusing multiple such side information estimates, generated with temporal and view interpolation methods, are proposed. Two algorithms are suggested along with weighted fusion schemes for combining multiple methods. The results presented indicate performance improvements of up to 2 dB compared to exiting fusion approaches.

Index Terms— DVC, Multiview, Fusion

1. INTRODUCTION

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.

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. Key frames are intra coded (in a lossy or lossless manner) with a conventional encoder and are made available at the decoder. WZ frame coding involves transformation and/or 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 is seen as the systematic part of the channel encoder's output as received at the decoder, i.e., the side information is seen as a noisy version of the original coded WZ frame. The received parity bits are used to correct the errors present in this noisy version of the WZ data. Clearly the quality of the side information will influence the performance of the DVC system [4] [5]. One of the most commonly used channel codes in DVC systems are the Rate Compatible Punctured Turbo (RCPT) codes [6] which allow for the amount of bits transmitted to be controlled by the receiver, assuming a feedback channel is present. In DVC the decoder will attempt to decode with a subset of the parity bits and will request for more only if decoding fails, with failure being defined as a symbol error rate above a specific threshold. More information on DVC can be found in [3], [4] and [5].

In this paper, we focus on the case of DVC applied to a multi-view scenario, where several cameras capture the same scene from different angles. In multi-view systems spatial/view correlations can be exploited for coding or view synthesis purposes [7], [8] and [9]. This paper develops novel techniques for fusing spatial and temporal estimates of the side information and is organised as follows. In section 2, existing multi-view DVC techniques are reviewed. Our approach is presented in section 3, with results following in section 4. Finally, section 5concludes this paper.

2. MUTLIVIEW DVC

The generation of the side information is an important part in the design of a DVC system as it directly relates to the system's rate distortion performance. The better the side information is, the fewer bits will be required for correctly decoding the WZ data. For single video source systems, motion interpolation using past and key frames is a common approach to forming the side information. However, in a multi source system, where different cameras capture the same scene from different angles, spatial correlation and redundancies between views can also be exploited for the generation of the side information. A classic structure for a multiview DVC system, using three cameras, can be found

in [10] and [11], where two cameras are intra i.e. they only generate key frames, with the third camera being a WZ camera, i.e. generating alternating WZ and Key frames. For the WZ camera, temporal motion interpolation (TMI) is available for generating a temporal estimate of the missing WZ frame. However, if scene and camera parameters are known, adjacent cameras can also be used to generate a spatial estimate which is computed using some form of spatial view interpolation (SVI). The side information can then be generated more reliably by merging these two estimates using fusion techniques. Typically these fusion techniques [10][11][12] will employ a binary mask, with 0 indicating pixels coming from the TMI estimate of the side information and 1 pixels predicted with SVI. Regions where TMI fails (e.g. due to high motion) should be estimated with SVI, whereas spatially occluded areas should be estimated with TMI. The generation of this binary mask dictates the quality of the side information and therefore drives the performance of the system.

A number of techniques for generating such a mask are proposed in [11]. Typically a prediction error will be estimated at the intra cameras, which will then be thresholded to generate a mask before the latter gets projected onto the WZ camera. Relevant to the above scenario is the method proposed in [13] for increasing the frame rate of a specific camera using an adjacent camera that operates at a higher frame rate. Previous and next frames from the two cameras are used in order to create disparity maps and generate predictors of the missing frames.

In both methods described above, it is assumed that adjacent cameras are fully-intra cameras. In [12] this is not the case anymore, with the structure consisting of N alternating WZ cameras, as shown in Figure 1, with key frames from each camera alternating both in time and space,. TMI is used only when the difference between the forward and backward motion compensated frames is smaller than a threshold T_1 and the estimated motion is smaller than a threshold T_2. Otherwise SVI is used. In this paper, this technique is referred to as *Motion Compensation Difference* (*MCD*). In [10], differences between the previous frame of the current WZ camera with corresponding TMI and SVI estimates are first computed. For each pixel, if the TMI estimate minimises the difference, TMI is associated with this pixel, otherwise SVI is used. The differences with the next frame of the current WZ camera are processed in a similar manner. These two masks are then merged into one by a logical OR operation. This algorithm is further combined with a scheme for restricting the pixels allocated to TMI according to the magnitude of the motion vectors. A simpler version consists of computing only the difference between the TMI estimate with the previous and next frames of the current WZ camera. These differences are thresholded to generate two binary masks which are merged via a logical OR operation. This simple technique is referred to as *Pixel Difference* (*PD*) in the remainder of the paper.

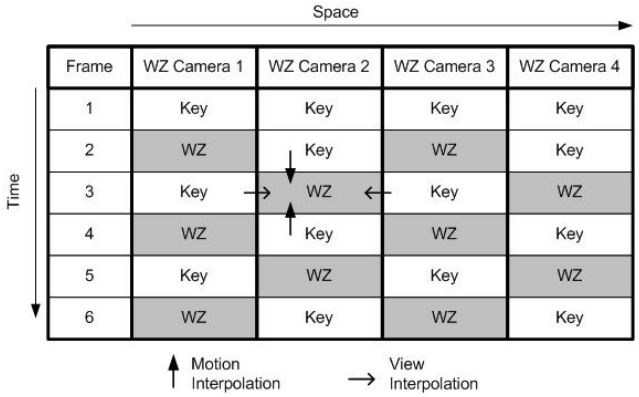


Figure 1: Multiview DVC structure under study

3. PROPOSED FUSION METHODS

The camera arrangement of the system under study is depicted in Figure 1. No purely-intra cameras are present making the fusion solutions of [11] and [13] unsuitable. With the proposed arrangement there are at least two key frames available at any time allowing depth maps to be generated for each of the key frames. The configuration of Figure 1 represents a flexible scheme which does not require the use of different types of cameras. The study presented in this paper could easily be extended to a higher number of WZ cameras. In the following, SVI is performed using view interpolation techniques of [7] and TMI is implemented using a bi-directional technique with spatial smoothing, as described in [5]. Novel algorithms have been implemented for fusing the two side information estimates (TMI and SVI).

3.1. Temporal Motion Interpolation Projection Fusion

In this method, the TMI estimate is projected onto the key cameras. Frame differencing is then applied between the original key frame of the key cameras (two in our case) and the projected TMI estimate. The frame differences are then thresholded in order to obtain two binary masks. For each pixel, if the difference is larger than a threshold, then TMI is assumed to have failed and the mask is set to 1. If the difference is smaller TMI is assumed to have performed well and the mask is set to 0. The two masks are projected back onto the current camera to form the final mask. This technique is referred to as *TMI projection*.

3.2. Spatial View Interpolation Projection Fusion

In this method, the SVI estimate coming from the key cameras is motion compensated, using the motion vectors of TMI, onto the previous and next key frames of the current camera. Frame differencing is then applied between these key frames and the motion compensated SVI estimate. The frame differences are then thresholded in order to obtain two binary masks. For each pixel, if the difference is larger than

a threshold, then SVI is assumed to have failed and the mask is set to 0. If the difference is smaller, SVI is assumed to have performed well and the mask is set to 1. The two masks are motion compensated, back onto the current frame to form the final mask. This technique is referred to as *SVI projection*.

3.3. Weighted Fusion through Reliability Masks

The two algorithms described above operate on binary masks, assigning hard coded pixel values to a particular estimate. When several masks are available it is possible to merge them using reliability levels [11]. This leads to the formation of some additional fusion methods. For the case of two masks, reliability levels can be assigned to each pixel based on the corresponding values of both masks. When the masks are in agreement (i.e. when the two masks have the same value) then the indicated method (TMI or SVI) has a reliability of 1. When the two masks are not in agreement then both methods share the same reliability value of 0.5 for the specific pixel and both contribute equally to the formation of this side information through a process of averaging. This merging process can be extended for the case of more than two masks with the reliability weights being adjusted according to the number of available masks and the times the use of a specific method is indicated.

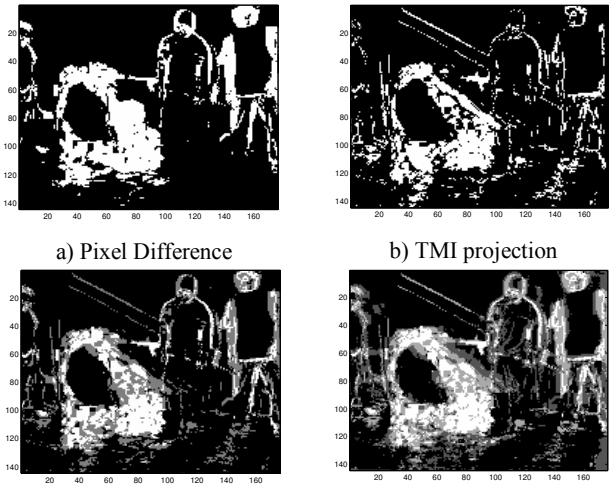
Figure 2 shows examples of binary masks for the PD and TMI projection algorithms and reliability masks for merging the results of the two methods (TMI and PD) as well as the case where three masks, resulting from TMI, PD and SVI respectively, are available. The grey levels indicate the contribution of the SVI estimate to the final side info pixel value, with lighter colours indicating a higher contribution.

4. RESULTS

Simulations were conducted using the *breakdancers* test sequence [14] at QCIF resolution (original resolution was 1024 x 768 - cropping followed by down-sampling was applied). Cameras 2, 3, 4 and 5 of the *breakdancers* setup have been used in our scenario (corresponding therefore to cameras 1, 2, 3 and 4 in Figure 1). After trial encoding, thresholds have been set to 10. First we study the performance of the binary mask schemes discussed so far, i.e. pure TMI, pure SVI, Motion Compensation Difference (MCD), Pixel Difference (PD), TMI projection and SVI projection. Figure 3 shows the PSNR values for the side information generated with these six methods. All values have been averaged over the whole sequence and over the four cameras. It can be seen that TMI projection generates the best side information. It would therefore be expected to provide the best rate distortion performance as well.

Figure 4 shows the rate distortion performance of the DVC system at 15fps when binary mask schemes are used. A comparison with H.264 Intra coding is also presented. All curves correspond to the average of the four cameras. Because of the very high motion activity TMI performs poorly whereas SVI does rather well. TMI projection offers the best results with an improvement of 0.6dB compared to the next best scheme. By projecting the TMI estimate to the key cameras the method can successfully identify the areas of high or irregular motion where pure TMI would normally fail to generate an accurate prediction. The use of SVI in these regions leads to better results. Compared to H.264 Intra, DVC offers better performance at low bit rates only.

The performance of a number of weighted fusion schemes was also examined and the results are shown in Figure 5. The methods examined were the following: TMI-projection/PD, SVI-projection/PD, TMI-projection/SVI-projection/PD and TMI-projection/SVI-projection/MCD. The rate distortion graphs of Figure 5 indicate that combining multiple methods through a weighted fusion scheme doesn't seem to offer any significant improvement, apart maybe at higher bit rates.



a) Pixel Difference b) TMI projection

c) TMI projection / PD d) TMI proj. / SVI proj. / PD

Figure 2: Example of binary and weighted fusion masks

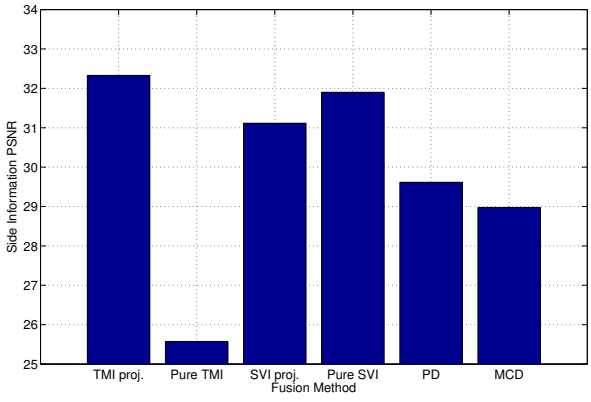


Figure 3: Side information PSNR with binary mask schemes

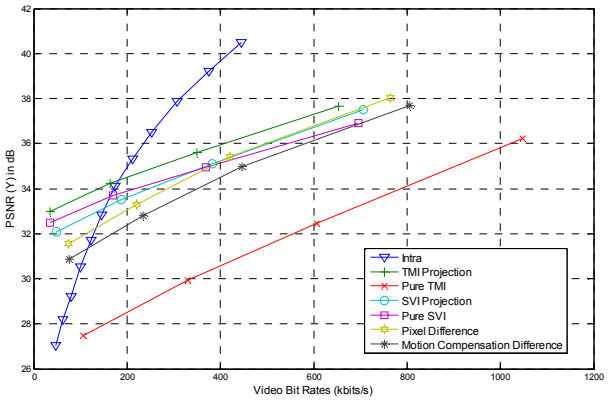


Figure 4: Rate/distortion at 15fps: Binary mask schemes

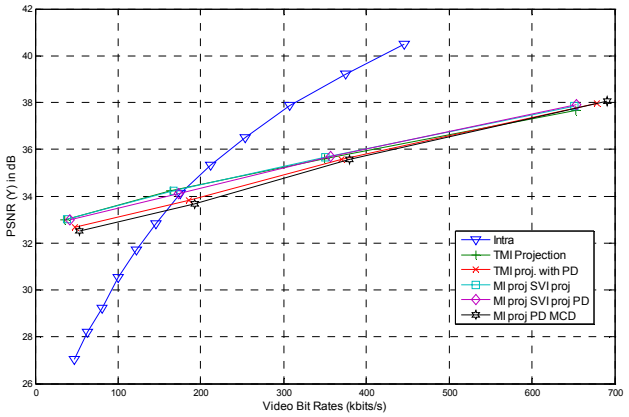


Figure 5: Rate/distortion at 15fps: Weighted fusion methods

5. CONCLUSION

This paper has studied and proposed algorithms for fusing side information estimates in multi-view DVC systems generated with temporal/motion (TMI) and spatial/view (SVI) interpolation methods. The camera configuration of the system examined in this work allows for a flexible multiview system consisting only of WZ cameras, with key frames alternating in space and time. The problem of side information generation under such a system configuration was studied and two solutions were suggested and tested for fusing the two side information estimates. The first one involves projection of the TMI estimates onto the adjacent cameras for the formation of binary masks which are then projected back onto the current camera. The second method included motion compensation of the SVI estimate. Results indicate that improvements of up to 2dB compared to competing methods can be obtained using the TMI projection approach. Weighted fusion schemes for combining multiple methods were also suggested and examined but were found to offer very little if anything in terms of performance gain.

6. ACKNOWLEDGMENT

The work was performed as part of the UK EPSRC VIGELANT project.

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]. D. Rowitch and L. Milstein, "On the Performance of Hybrid FEC/ARQ Systems Using Rate Compatible Punctured Turbo (RCPT) Codes," IEEE Trans. on Communications, , no. 6, pp. 948 – 959, 2000.

[7]. S.E. Chen and L. Williams, "View interpolation for image synthesis," Computer Graphics, vol. 27, 1993.

[8]. H.Y Shum and S. B. Kang, "A review of image-based rendering techniques," in VCIP, Perth, June 2000.

[9]. R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, Second Edition, Cambridge, 2003.

[10]. M. Ouaret, F. Dufaux, and T. Ebrahimi, "Fusion-based Multiview Distributed Video Coding," in IEEE ACM VSSN, Santa Barbara, October 2006.

[11]. X. Artigas, E. Angeli, and L. Torres, "Side Information Generation for Multiple Distributed Video Coding Using a Fusion Approach," in - NORSIG, Reykjavik, June 2006.

[12]. X. Guo, Y. Lu, F. Wu, W. Gao, and S. Li, "Distributed multi-view video coding," in SPIE, San Jose, January 2006.

[13]. D. Hazen, R. Puri, and K. Ramchandran, "Multi-camera video resolution enhancement by fusion of spatial disparity and temporal motion fields," in IEEE International Conf. on Computer Vision Systems, New-York, 2006, p. 38.

[14]. C.L Zitnick, S.B Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in ACM SIGGRAPH, Los Angeles, August 2004, pp. 600–608.