

Side Information Generation for Multiview Distributed Video Coding Using a Fusion Approach

Xavi Artigas, Egon Angeli, and Luis Torres

Technical University of Catalonia (UPC), Department of Signal Theory & Communications
Campus Nord, D-5. Jordi Girona 1-3, 08034 Barcelona, SPAIN
E-mail: {xavi, aleston, luis}@gps.tsc.upc.edu

ABSTRACT

Distributed Source Coding (DSC) aims at achieving efficient compression by locating the source redundancies at the decoder instead of the encoder. Moreover, DSC exhibits many properties like low-complexity encoding or embedded error resilience that make it very convenient for some emerging new applications. Among the many challenging topics related to DSC there is the generation of the Side Information, an estimation made by the decoder of the data being decoded. In the particular field of Multiview Distributed Video Coding (Multiview DVC) this Side Information can be generated by inter-camera or intra-camera interpolation. This paper briefly describes both techniques and proposes two approaches that combine them by evaluating the reliability of each interpolation at the pixel level.

1. INTRODUCTION

Video coding research and standardization have been adopting until now a video coding paradigm where it is the task of the encoder to explore the source statistics, leading to a complexity balance where complex encoders interact with simpler decoders. Distributed Video Coding (a particularization of Distributed Source Coding) adopts a completely different coding paradigm by giving the decoder the task to exploit the source statistics to achieve efficient compression. This coding paradigm is particularly adequate to emerging applications such as wireless video cameras and wireless low-power surveillance networks, disposable video cameras, medical applications, sensor networks, multi-view image acquisition, networked camcorders, etc., where low complexity encoders are a must because memory, computation, and energy are scarce.

However, even though the theoretical bases for Distributed Source Coding were set thirty years ago with the work by Slepian & Wolf [1] (for the lossless case) and Wyner & Ziv [2] (for the lossy case), it has been only recently that research on the topic has taken a new momentum. This research has been encouraged by the rise of some new applications, and has been led mainly by Ramchandran *et al.* [3] and Girod *et al.* [4]. A good review of other works can be found in [4].

On the other hand, Multiview techniques have been researched in the past, both for coding [5] and for camera interpolation, since they allow creating views from virtual (non-existent) cameras, or what is called Free Viewpoint Navigation of scenes given only recordings from a few cameras [6].

The objective of Multiview DVC is to efficiently encode different video streams, but exploiting the possible redundancies at the decoder, thus obtaining benefits inherent to DVC like lower encoding complexity, embedded error resilience or the fact that no connection is necessary between the different cameras. This paper presents a method to apply Distributed Video Coding concepts to the Multiview problem.

Multiview DVC has only recently received attention from the scientific community. Ramchandran *et al.* [7], Girod *et al.* [8] and Guo *et al.* [9] have published some of the work dealing with this topic. [7] and [8] work with static images. The technique described in [9] is summarized later in section 3 and compared to the proposed methods.

1.1 Problem statement

The following minimal setup is proposed. It can be further augmented by adding more cameras or changing their configuration, but this is the strictly minimum structure to describe the proposed techniques. Three cameras are used, which do not communicate among them as stated by the DSC theorems. Two of them are called **Intra Cameras** and work in a conventional fashion, i.e., their video stream is encoded and decoded independently of the other cameras. The third camera, called **Wyner-Ziv (WZ) camera**, independently encodes but requires the video streams from the other cameras for decoding (Fig. 1). This joint decoding allows the WZ camera to transmit at a lower rate than if it was decoded on its own, as stated by the Slepian-Wolf theorem.

The Wyner-Ziv camera transmits some frames in Intra mode, as in [4]; this is, coded independently of the other frames. The rest of the frames are called Wyner-Ziv frames and are the ones that will benefit from the joint decoding performed at the receiver (Fig. 2). In the following explanations it will be assumed that only one WZ frame is present between every two Intra frames.

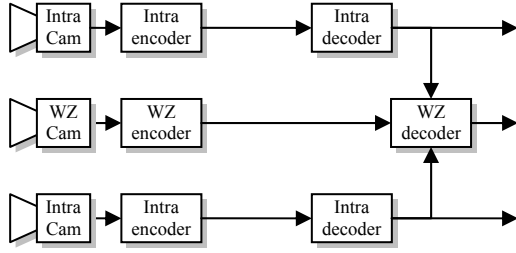


Fig. 1. General setup. Intra cameras operate in a conventional fashion while the Wyner-Ziv camera requires joint decoding.

To decode a WZ frame the decoder first needs to generate the side information, which is an estimation for the frame. The better the side information (the more correlated it is with the frame being estimated), the fewer bits will be required to encode the WZ frame. The proposed methods deal with the generation of the side information through interpolation.

2. INTRA/INTER CAMERA INTERPOLATION

As can be seen in Fig. 2, a WZ frame has a number of nearby frames that can be used to generate its side information. The following two subsections describe two different interpolation schemes that can be used to this avail. As it turns out, these interpolations perform better when combined, and for that reason, the next section presents different mechanisms to combine them.

2.1 Intra-Camera Interpolation

The method that uses only information from the WZ camera to estimate a WZ frame is called Intra-Camera Interpolation (IntraCI), or Temporal Interpolation. Its goal is to generate the frame in between two given Intra frames ($k-1$ and $k+1$) that resembles the most the original frame k , which is not available. A technique commonly used to carry out this task is Motion Compensated Temporal Interpolation (MCTI) [10], which performs Block-based Motion Estimation between the two Intra frames, and creates the estimated frame by using halved motion vectors. MCTI has also been used, for example, to perform temporal up-sampling [10].

It is worth noting that the fact that the frame being estimated is not available makes this method slightly different to conventional Motion Compensated Prediction (MCP) as used in hybrid coding. In MCP motion vectors are computed between frames $k-1$ and k , while in MCTI motion vectors are computed between frames $k-1$ and $k+1$, and used to estimate frame k .

MCTI tends to fail when there is rapid movement, or when the movement does not follow the translational model assumed by the algorithm. In the DVC setup, this problem can be solved by using information from the other cameras, by means of the technique described in the next subsection.

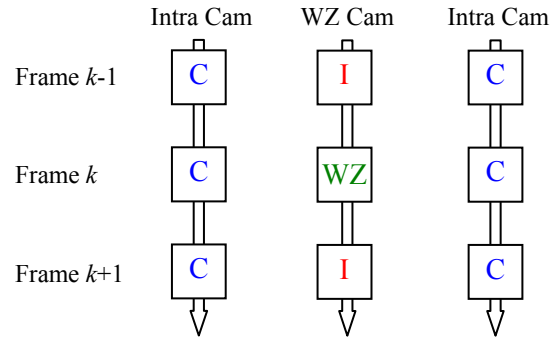


Fig. 2. Spatiotemporal frame structure. Frames labeled “C” belong to a conventionally encoded video sequence. “I” frames are encoded as single images. “WZ” frames use the distributed coding scheme.

2.2 Inter-Camera Interpolation

When only information from other cameras is used to estimate the frame at the WZ camera, the method is called here Inter-Camera Interpolation (InterCI), or Spatial Interpolation. These methods are also known as Image Based Rendering (IBR) and have been intensively researched in the past [6].

The algorithm chosen for this work is similar to View Interpolation (VI) [11] and requires that a depth map is available for each frame of the Intra cameras that is to be used to estimate the WZ camera. Depth maps specify the distance of each pixel in an image to the camera that took the image, and can be calculated using computer vision techniques given two pictures from two calibrated cameras (this is the reason why two Intra cameras are required in this setup). A calibrated camera is a camera whose parameters (position and orientation in the scene, focal length ...) are known. Many other algorithms exist to perform the Inter-Camera Interpolation [6].

View Interpolation works as follows. Given a frame from an Intra camera, its associated depth map, and the known parameters of the Intra camera, every pixel in the frame is projected back to the scene, i.e., the 3D coordinates of the real-world object that originated that pixel are calculated. This process creates a point cloud that encapsulates all the information that the Intra frame contains about the scene. Then, given the known parameters of the WZ camera, every point in the cloud is projected onto the WZ camera, thus producing the estimated frame.

One drawback of VI is that those regions of the scene for which the Intra camera has no information (due to *occlusions*) appear completely black in the WZ frame. If this information is not added to the process by other means, only error concealment techniques can alleviate this problem. For the non-occluded areas InterCI also introduces other kinds of errors due to the depth maps not matching the real depth, and view-dependant scene features like reflections, that change when the camera changes, and therefore cannot be easily interpolated.

3. INTRA/INTER CAMERA FUSION

The two interpolation techniques presented above estimate correctly some parts of the WZ frame, but fail in other parts: IntraCI has problems with high motion areas and InterCI cannot easily deal with scene occlusions and reflections. To overcome these difficulties a mechanism is designed that fuses the correctly predicted parts of each estimation and discards the others. Such a fusion mechanism requires the creation of a Reliability Measure that indicates which pixels in an interpolated frame have potentially been correctly estimated. Three different reliability measures are presented in the following subsections which originate three different fusion algorithms; one already present in the literature, and two novel ones.

3.1 Estimation of the Motion Compensation Error

MCTI is used on Intra frames of the WZ camera to generate a temporal interpolation, which is based on block matching. The block in frame $k+1$ that is most similar (according to some metric) to a given block in frame $k-1$ is said to be that same block, after translation. If the difference between the original block in frame $k-1$ and the final candidate in frame $k+1$ is low, they are probably the same block, but if this difference is high, the MCTI has probably chosen a bad candidate. Therefore this difference can be used as a reliability measure.

The Estimation of the Motion Compensation Error (EMCE) technique thresholds this difference to obtain a binary reliability mask. IntraCI is used for those pixels where MCTI is reliable and InterCI is used for the rest. For those pixels for which InterCI provides no information (due to occlusions), MCTI is used (even when it was deemed unreliable). This technique has already been described in [9] with an additional threshold on the length of the motion vectors.

One problem of this technique is that the threshold on the error measure is highly sequence-dependant (and even frame-dependant). One second, and worse, shortcoming is that sometimes MCTI finds a perfect match between two blocks, resulting in extremely high reliability, but the generated frame is not correct due to a missed moving object. When this happens, information from other cameras regarding the missed object will not be used due to the high reliability assigned to the MCTI estimation.

3.2 Projection of the Motion Compensation Error

Instead of trying to estimate the MCTI reliability as EMCE does, the exact MCTI error can be calculated if the original frame is available. It is not available for the WZ camera, but it is for the Intra cameras. This is the idea behind the Projection of the Motion Compensation Error (PMCE) technique proposed in this work.

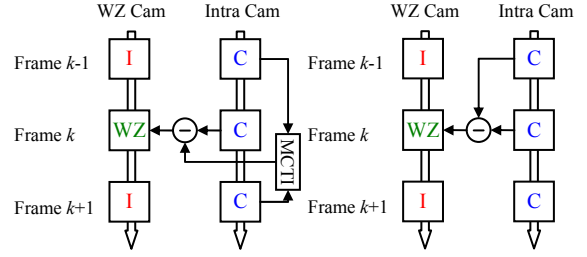


Fig. 3. Structure of the PMCE and PNFE techniques.

The process is depicted in Fig. 3-left. An MCTI is performed between frames $k-1$ and $k+1$ of an intra camera. The result is subtracted from frame k (the frame being estimated) providing the exact prediction error. This error is again thresholded to obtain a binary reliability mask, but this threshold is far less sequence-dependant than the one used in EMCE. This mask must now be projected onto the WZ camera using InterCI. After this step, the mask no longer indicates the exact MCTI error for two reasons: First, InterCI is not an ideal process (it requires ideal depth maps) and introduces errors of its own, and second, the MCTI performed on the WZ camera does not need to fail in the exact same spots as the MCTI performed on the Intra camera. For these reasons, the obtained mask is not anymore an error indicator but a reliability indicator.

Once the mask on the WZ camera is obtained the algorithm proceeds as in EMCE: areas where the MCTI is marked as reliable use IntraCI, and the rest use InterCI.

3.3 Projection of the Neighboring Frame Error

Since MCTI will produce different results when executed on different cameras, the error in one camera will not generally be adequate as a reliability indicator for the other one. Projection of the Neighboring Frame Error (PNFE) aims at solving this issue and is the second technique proposed in this work.

As the previous technique, PNFE makes a prediction for frame k at an Intra camera, and compares the prediction with the original frame. In this case, the prediction is simply one neighboring frame (either $k-1$ or $k+1$). As shown in Fig. 3-right, both frames are subtracted and thresholded, to obtain the binary reliability mask, and then projected onto the WZ camera using InterCI.

To generate the estimation for the WZ frame, the algorithm has two predictors from the same WZ camera: the previous and the next frame. For each one of these predictors, there are as many reliability masks as Intra cameras. The first step is to fuse the masks from all Intra cameras, and this is done by a simple logical OR operation: a pixel is marked as unreliable if any Intra camera thinks that the pixel is unreliable.

At this point, each predictor has an associated reliability mask, and the final estimation is generated following

these simple rules: For each pixel, if both predictors are reliable, they are averaged. If only one is reliable, it is directly used, and if none of the predictors is reliable, the pixel is filled using InterCI.

4. EXPERIMENTAL RESULTS

The Ballet and Breakdancers test sequences [12] (1024x768 @ 15Hz., 100 frames) have been used in the simulations. Cameras 0 and 2 have been used uncoded as Intra cameras, and camera 1 has been the WZ camera. Even frames of camera 1 have been used uncoded as Intra frames, and odd frames have been estimated (Side Information in the DVC context) and compared to the original ones to obtain the PSNR figures. The thresholds for EMCE are the optimal for each sequence. The average PSNR quality of the estimated frames can be seen in Fig. 4 and Fig. 5 for each studied technique. It can be seen in Fig. 4 that IntraCI works better than InterCI for this sequence. This is mainly due to occlusions and reflections. Conversely, in Fig. 5, it is shown that InterCI works significantly better than IntraCI. This is due to high motion and low temporal sampling rate. It can also be observed that the proposed fusion techniques work better than IntraCI or InterCI on their own.

As an example, Fig. 6 shows a portion of the obtained reliability masks for the three techniques. White areas indicate that IntraCI is unreliable and thus InterCI is used. It can be seen that the EMCE mask works on a block level and that MCTI is a more reliable predictor (fewer white pixels) than neighboring frames. The superior performance of PNFE over PMCE is explained because the PMCE mask is itself unreliable, since it assumes that MCTI will work the same way on different cameras.

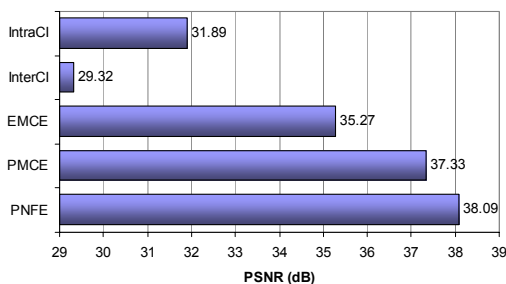


Fig. 4. PSNR of the estimated frames for “Ballet”.

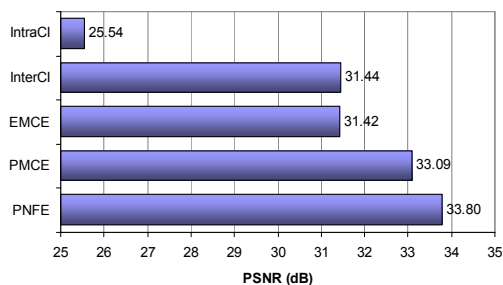


Fig. 5. PSNR of the estimated frames for “Breakdancers”.



Fig. 6. Reliability masks for a)EMCE, b)PMCE, c)PNFE

5. CONCLUSIONS

Two techniques have been presented to generate side information in the context of Multiview Distributed Video Coding. Results show that the Projection of the Neighboring Frame Error technique provides the side information that resembles the most the original frame being estimated. Since the side information available at the WZ decoder is better, the WZ encoder should need to transmit fewer bits when this technique is used. Research is on the way to validate this assessment.

ACKNOWLEDGMENT

The work presented was developed within DISCOVER, a European Project (www.discoverdvc.org), funded under the European Commission IST FP6 programme, and by grant TEC2005-07751-C02-02 of the Spanish Government.

REFERENCES

- [1] D. Slepian and J. Wolf, “Noiseless coding of correlated information sources”, *IEEE Trans. Inform. Theory*, vol. 19 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. Inform. Theory*, vol. 22, pp. 1-10, January 1976.
- [3] R. Puri and K. Ramchandran. “PRISM: A new robust video coding architecture based on distributed compression principles”. *Proc. of 40th Allerton Conf. on Comm., Control, and Computing*, Allerton, IL, Oct. 2002.
- [4] B. Girod, A. Aaron, S. Rane and D. Rebollo-Monedero, “Distributed video coding”, *Proc. of the IEEE*, vol. 93, no. 1, January 2005
- [5] Jens-Rainer Ohm, “Stereo/Multiview Video Encoding Using the MPEG Family of Standards”, Invited Paper, Electronic Imaging '99, San Diego, Jan. 1999
- [6] Heung-Yeung Shum and Sing Bing Kang. "A Review of Image-based Rendering Techniques", *IEEE/SPIE Visual Communications and Image Processing (VCIP) 2000*, pp. 2-13, Perth, June 2000
- [7] G. Toffetti, M. Tagliasacchi, M. Marcon, A. Sarti, S. Tubaro, K. Ramchandran, “Image Compression in a Multi-camera System based on a Distributed Source Coding Approach”. *European Signal Processing Conference*, Antalya, September 2005
- [8] X. Zhu, A. Aaron and B. Girod, “Distributed compression for large camera arrays”, *Proc. IEEE Workshop on Statistical Signal Processing, SSP-2003*, St Louis, Missouri, Sept. 2003.
- [9] X. Guo, Y. Lu, F. Wu, W. Gao, S. Li, “Distributed multiview video coding”, *Proceedings of SPIE*, January 2006, San Jose, California, USA. Vol. #6077.
- [10] S.H. Lee, O. Kwon, R. H. Park, “Weighted-Adaptive Motion-Compensated Frame Rate Up-Conversion”, *IEEE Trans. on Consumer Electronics*, Vol. 49, No. 3, 2003.
- [11] S. E. Chen and L. Williams, “View interpolation for image synthesis”. *Computer Graphics*, 27: 279-288, 1993.
- [12] <http://research.microsoft.com/vision/InteractiveVisualMediaGroup/3DVideoDownload/>