Stereo side information generation in low-delay distributed stereo video coding

Salmistraro, Matteo; Forchhammer, Søren

Published in:
Proceedings of SPIE, the International Society for Optical Engineering

Link to article, DOI:
10.1117/12.929230

Publication date:
2012

Stereo side information generation in low-delay distributed stereo video coding

Matteo Salmistraro, Søren Forchhammer

Department of Photonics Engineering, Technical University of Denmark, Ørsteds Plads, 2800 Kgs. Lyngby, Denmark

ABSTRACT

Distributed Video Coding (DVC) is a technique that allows shifting the computational complexity from the encoder to the decoder. One of the core elements of the decoder is the creation of the Side Information (SI), which is a hypothesis of what the to-be-decoded frame looks like. Much work on DVC has been carried out: often the decoder can use future and past frames in order to obtain the SI exploiting the time redundancy. Other work has addressed a Multiview scenario; exploiting the frames coming from cameras close to the one we are decoding (usually a left and right camera) it is possible to create SI exploiting the inter-view spatial redundancy. A careful fusion of the two SI should be done in order to use the best part of each SI. In this work we study a Stereo Low-Delay scenario using only two views. Due to the delay constraint we use only past frames of the sequence we are decoding and past and present frames of the other. This is done by using Extrapolation, to exploit the time redundancy and well known techniques for stereo error concealment. This allows us to create good quality SI even if we are only using two views. In this work we have also used a new method in order to fuse the two SIs, inspired by Multi-Hypothesis decoding. In this work the multiple hypotheses are used to fuse the SIs. Preliminary results show improvements up to 1 dB.

Keywords: Distributed Video Coding, DVC, Multiview DVC, Stereo DVC, Extrapolation, Difference Projection.

1. INTRODUCTION

In recent years Distributed Video Coding (DVC) gained interest as research topic, due to the possibility of developing low-complexity video encoders. In fact, it is possible to leverage the time redundancy of a video sequence at the decoder and not at the encoder, shifting the complexity to the decoder and allowing the encoder to process the frames independently. A DVC[1,2] based system could be a good solution in the case of a large amount of video sources which send video streams to an unique decoder (e.g. in video surveillance). In this scenario a low-complexity and maybe efficient encoder is desired and possibly the complexity constraint is more relaxed at the decoder side.

DVC is based on two information theory theorems, the Slepian-Wolf[3] and the Wyner-Ziv[4] theorems, where, in the second case, source data are independently lossy coded but jointly decoded using a correlated source at the decoder, this correlated source is commonly called Side Information (SI).

In single view DVC, one or more SI are used, they are generated exploiting the temporal redundancy. In Multiview DVC (M-DVC) there are more than one video stream, these video streams are correlated, i.e. they show the same scene from different points of view. This allows the system to exploit also the spatial redundancy, which provides the possibility of predicting part of an unknown frame using other views. Usually this prediction (spatial prediction) is less accurate compared to the temporal one but they can be used together in order to improve the performance of the system.

A decisive advantage of M-DVC compared to standard multiview coding[5] is that it does not require inter-camera communication in order to exploit spatial redundancy, because the spatial redundancy is exploited at the decoder, allowing the sources to be totally independent.

Further author information:
Matteo Salmistraro: E-mail: matsl@fotonik.dtu.dk, Telephone: +45 45256635

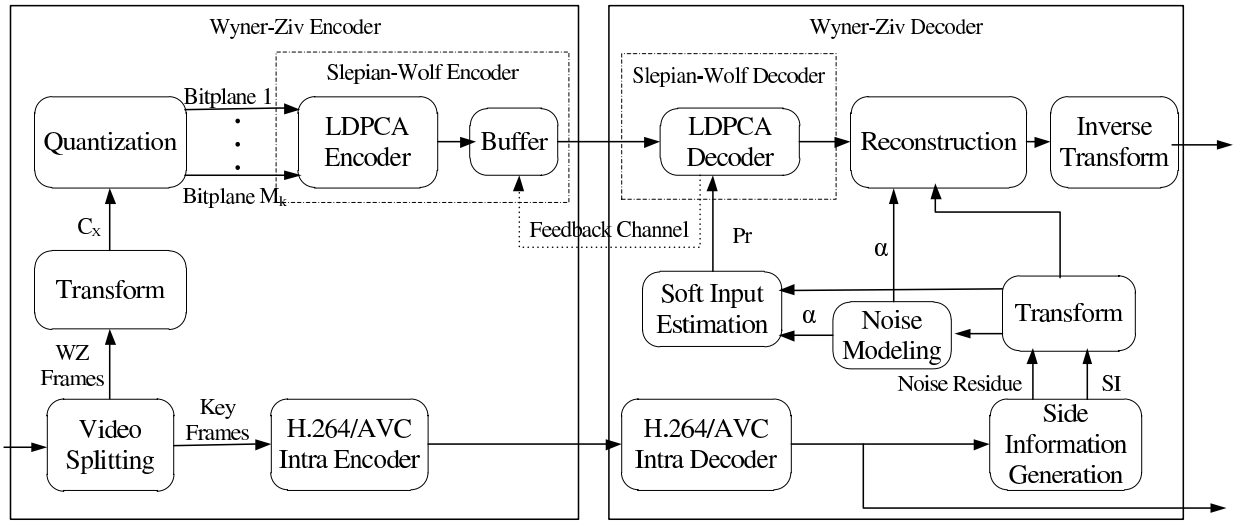

Figure 1. The feedback-based DVC system[9].

In this paper the stereo scenario is addressed. In this case there are only two views to code. They are usually used to give the illusion of depth to the viewer, hence while decoding one view we can only use the other one for leveraging the spatial redundancy. On the other hand the stereo video has particular geometrical constraints, which can be used to have good quality SI even if only two views are used.

We focussed on a low-delay scenario, in which future frames cannot be used in order to predict the current one, but the current frame in the other view can be exploited. Extrapolation[6] is used in order to produce the temporal SI, difference projection[7] and the similarity of the motion vector between views[5] will be also used in order to produce spatial SIs.

The contributions of this paper are: the study of stereo M-DVC in the low-delay case and the assessment of the quality level of the SI using stereo sequences, instead of standard sequences[8]. Finally the various SIs generated will be fused using a multi-hypothesis approach[6].

This paper is organized as follow: in Section 2 the DVC decoder architecture is presented. The methods used for producing the SIs are introduced in Section 3. Section 4 discusses the results.

2. TRANSFORM DOMAIN WYNER-ZIV VIDEO CODING

The state-of-the-art Transform Domain Wyner-Ziv (TDWZ) DVC decoder used in this paper was presented in[9]. In this system the frames are divided into key frames and Wyner-Ziv (WZ) frames. The first ones are compressed using conventional video compression systems like H.264/AVC intra coding. The WZ frames are transformed using a 4x4 DCT, quantized and decomposed into bitplanes. Each bitplane is coded using a Rate-Adaptive LDPC Accumulate (LDPCA) encoder[10], only a subset of the parity bits are sent to the decoder and the systematic part is discarded. The decoder uses the corresponding bitplane obtained from the SI as systematic part and corrects the differences between the SI and the original bitplane with the aid of the parity bits. The DCT coefficients are decoded following the zig-zag scan order and for each coefficient the MSB (Most Significant Bitplane) is decoded first, while the LSB (Less Significant Bitplane) is decoded last. If the decoder is not able to obtain an acceptable solution new parity bits are requested through the feedback-channel. In Figure 1 the architecture of the system is depicted.

The SI is the prediction made at the decoder of the frame which have to be decoded. This prediction can be made leveraging the spatial or temporal redundancy. The SI is only available at the decoder.

The LDPCA decoder requires the conditional probabilities $p(X_i = 0|Y)$ where X_i is the i-th bit of the bitplane which has to be decoded and Y is the Side Information. In order to calculate these probabilities it is

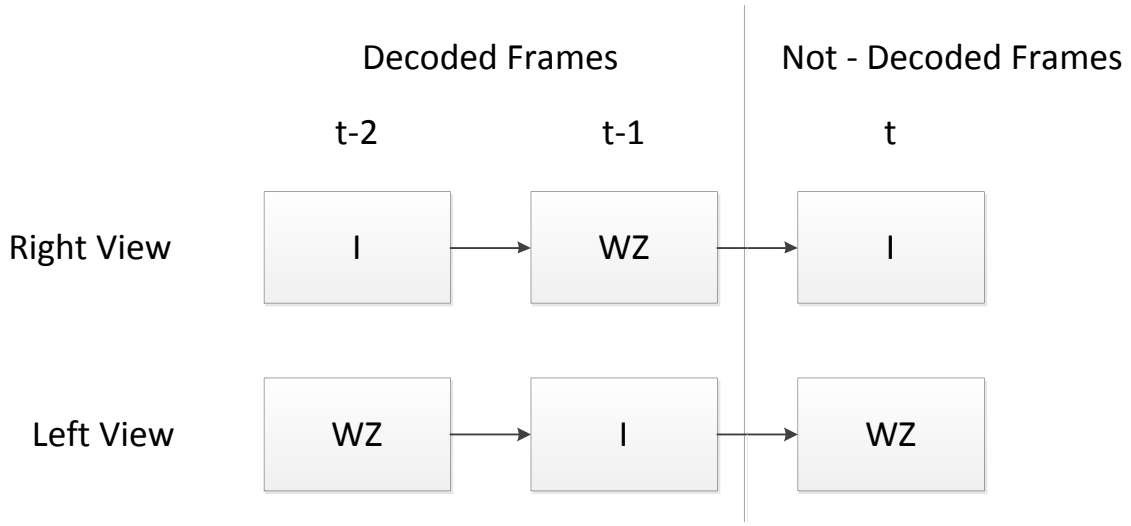

Figure 2. A possible video stream structure compatible with our approach.

Figure 3. Extrapolation Algorithm.

also necessary to have a noise model in order to model the difference (noise) between the original frame (X) and the SI. After the decoding of the bitplanes the frame is reconstructed obtaining the decoded version $\hat{X}$.

3. SIDE INFORMATION GENERATION

A low-delay scenario is addressed: future frames cannot be used in order to produce the SI for the current one. In this paper the frame to be decoded is the one at instant t, we will assume that all the frames at instants $t-i$, $i>0$ have been received and correctly decoded in both views and that the frame at instant t in the other view is also available. This is not a limiting assumption, since in order to obtain this it is sufficient that the sequences of key frames and WZ frame are shifted of one position with respect to each other in the two views, see Figure 2.

3.1 Extrapolation

Extrapolation has been discussed in a number of publications[6]. In the DVC architecture introduced in Section 2 the encoder can access past and future frames in order to produce the temporal SI. We will refer to this SI as interpolation, it has also been called differently e.g. MCTI. The interpolation allows the system to obtain good performance but on the other hand this increases its delay, for further details please refer to[6]. Extrapolation relays only on past frames in order to predict the current one, hence it achieves lower delay but also lower performance due to the inability of the system to adapt to sudden motion changes. The extrapolation module presented in[6] has been used in this work. In order to make this paper self-contained we will briefly summarize the method in this subsection. The structure of the algorithm is depicted in Figure 3, with this procedure we estimate the frame at instant t knowing the frames at instants $t-1$ and $t-2$. The first step is the motion estimation: via an 8x8 pixels block matching procedure the motion vectors between the frames at $t-2$ and $t-1$ are estimated. The Mean Square Error (MSE) is used as measure of the reliability of the Motion Vector (MV) the higher the MSE the lower the reliability. The MV field is smoothed using the 8 closest neighbor via a weighted average, the higher the reliability the higher the weight in the average. The next step is the refinement:

an 8x8 block size is a good compromise between flexibility and robustness but in some areas of the image the blocks could not be small enough to capture the movement. If the MSE is too high the corresponding blocks are divided into 4x4 pixels blocks, the old MV is used as initial estimation and then refined using a 4x4 pixels block matching. Finally the estimated motion field is applied to the frame at $t-1$ in order to estimate the frame at instant t, in the case of overlapping areas between blocks the values belonging to the more reliable block is chosen to fill the common pixels. Finally the unfilled pixels (holes) in the generated frame are filled using the MV belonging to the closest block. After this last step the SI generated by the extrapolation module Y_{Ex} is ready for being used in the decoding.

3.2 Difference Projection

In usual multiview video streams we cannot assume some a-priori knowledge of the cameras' placement if we are not targeting specific scenarios, but in the stereo case we can make these assumptions without loss of generality. For stereo streams various studies have been carried out for error concealment[11], developing techniques allowing the receiver to manage lost packets using the decoded ones. We focused on full frame recovery techniques[7], since the problem of full frame recovery is quite similar to the SI generation problem, in both cases an estimation of an unknown frame must be created using other frames.

The main idea behind the difference projection technique[7] can be summarized as follows: suppose, for the sake of argument, that the right frame at instant t have to be predicted, knowing the right frame at instant $t-1$ and the left frames at instants t and $t-1$. We can focus our attention on a point $p_{(l,t)}$ belonging to the left frame at instant t, its co-located point in the past frame is $p_{(l,t-1)}$. The disparity field between the left and the right frames at instant $t-1$ can be estimated obtaining $p_{(r,t-1)}$ which is the point corresponding to $p_{(l,t-1)}$. Denoting with $I(p)$ the intensity value of an image in the point p the following expression can be written:

$$\delta_l = I_{l,t-1}(p_{(l,t-1)}) - I_{l,t}(p_{l,t}) = I_{(l,t-1)}(p_{(l,t)}) - I_{(l,t)}(p_{(l,t)}) \tag{1}$$

the assumption that $\delta_r \approx \delta_l$ is quite reasonable, thus the value of $I_{(r,t)}(p_{(r,t)})$ can be estimated:

$$I_{(r,t)}(p_{(r,t)}) = Y_{dp}(p_{(r,t)}) = I_{r,t-1}(p_{(r,t-1)}) - \delta_r \approx I_{r,t-1}(p_{(r,t-1)}) - \delta_l \tag{2}$$

Obviously this process could create artifacts, hence this method is only applied to the pixels which experience high intensity differences. If the intensity difference is low the co-located pixel in the right frame at instant $t-1$ is copied. This process is done defining the change detection map $M_{(l,t-1\to t)}$. In the original formulation this could lead to project (apply the algorithm to) isolated pixels or not to project a particular pixel surrounded by projected pixels. The original method has been developed for maximizing the PSNR, now it is also important to avoid creating artifacts which could modify the distribution of the high-frequency DCT coefficients. In order to avoid this problem the change detection map is post-processed before using it to decide whether or not a pixel should be projected. Isolated pixels are deleted and pixels surrounded by other projected pixels are also projected.

3.3 Motion Vector Similarity

The motion vector similarity has been widely used in M-DVC[5], using the notation and the hypothesis in the past subsection, the motion vectors between $I_{(l,t)}$ and $I_{(l,t-1)}$ can be estimated, starting from the point $p_{(l,t-1)}$ the corresponding point in the next frame $p'_{(l,t)}$ can be found.

$$mv_{t,t-1} = p_{(l,t)} - p'_{(l,t-1)} \tag{3}$$

using the disparity estimation also used in the past section, $p_{(r,t-1)}$ can be identified and $p_{(r,t)}$ can be calculated

$$p_{(r,t)} = mv_{t,t-1} + p_{(r,t-1)} \tag{4}$$

which is used to project $I_{(r,t-1)}$ to Y_{mvS}

$$Y_{mvS}(p_{(r,t)}) = I_{(r,t-1)}(p_{(r,t-1)}) \tag{5}$$

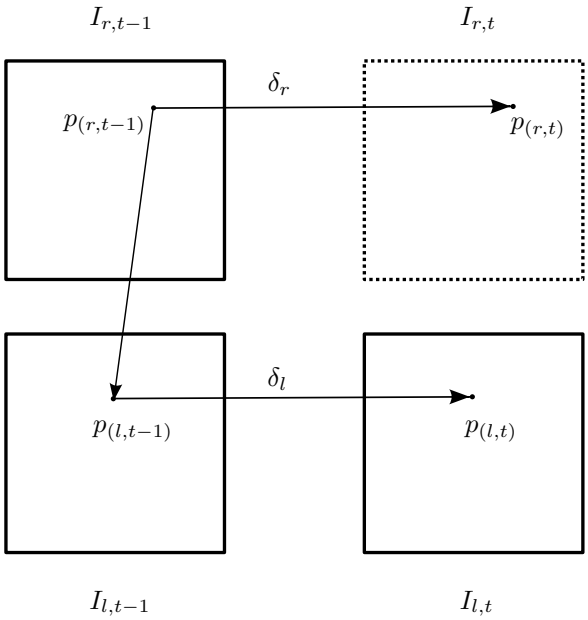

Figure 4. Difference Projection Method.

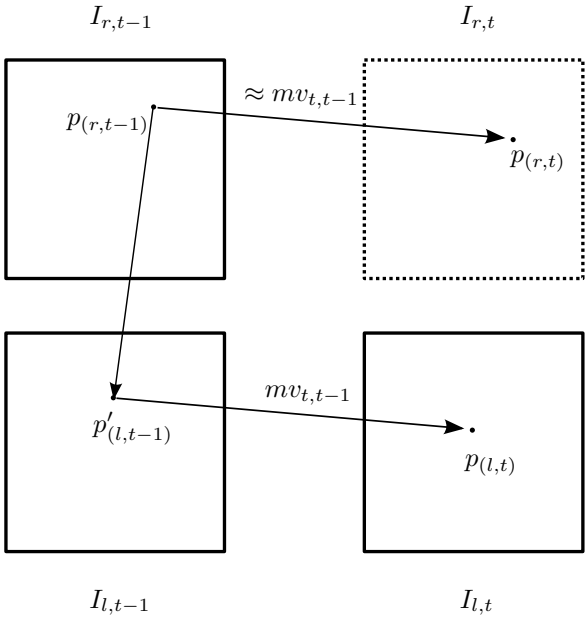

Figure 5. Motion vector similarity method.

3.4 Multi-Hypothesis Decoder

In[12] the authors presented a multi-hypothesis decoder. In a multi-hypothesis decoder there are more than one SI. For every SI the corresponding bitplane is extracted and it is feed into the LDPCA decoder (see Fig. 6), the first decoder which is able to decode successfully is the winning decoder for that bitplane and its output is taken as the decoded string after confirmation by an 8 bit CRC. In the presented system there are three different decoders, one for every generated SI. The winning decoder defines also the SI used for the reconstruction. This method has been chosen, instead of more conventional block-fusion systems[8] since in this case all the three SIs have comparable average quality levels, secondly the fusion between SI is a complex problem to solve, with a

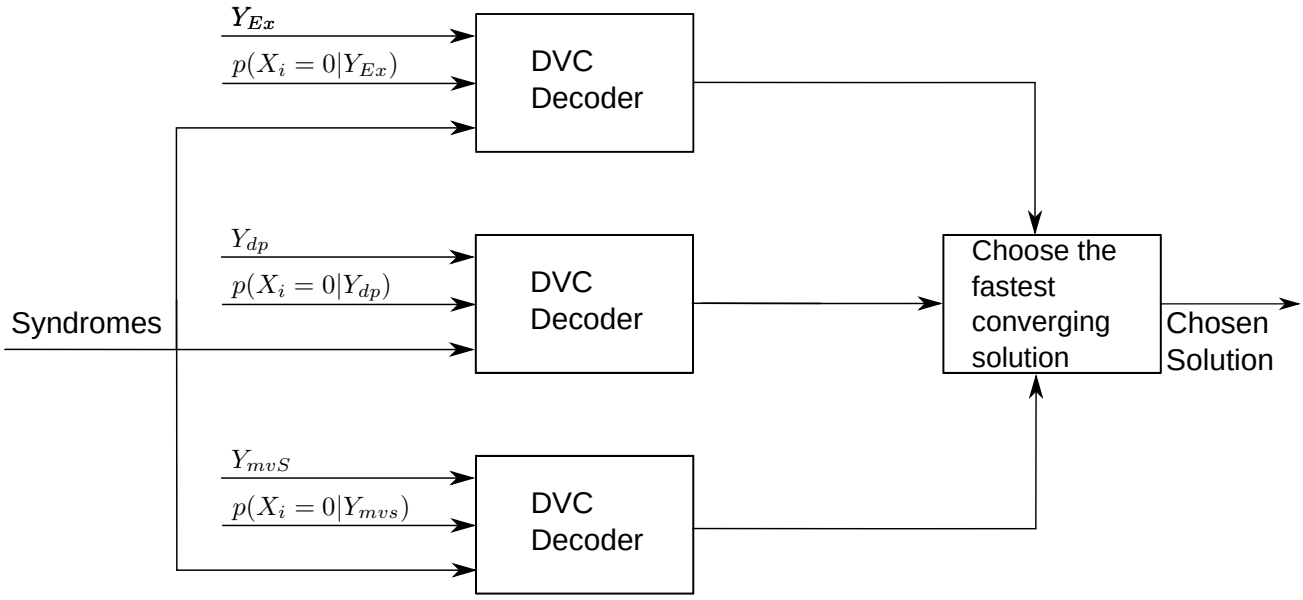

Figure 6. Multi-Hyphothesis Decoder.

multi-hypothesis decoder the problem is less complex, because various SIs, produced in different manners, could be fused in various ways automatically adapting the fusion strategy to the particular frame. In this paper the SIs will not be fused using different weights as in[12] but simply the one requiring less bits will be used on a bitplane-basis. The motivation is that an off-line residual is used in this paper. The usual SI fusion[12] uses the estimated residual in order to fuse only the unreliable parts of the SI, using this technique in this case could lead to incorrect conclusions.

4. RESULTS

In this section the performance of the system is evaluated. The LDCPA code[10] used a block length of 4800 bits. A residual estimation system has not been developed, instead for the experiments an off-line residual has been used, i.e. the residual R is calculated as $R = Y - X$, where X is the original frame and Y the SI used in the given decoder. In order to make a fair comparison also the extrapolation results are produced using the off-line residual. It has to be stressed that in this work we want to assess the quality of the SIs and the possible use of a Multi-Hypothesis decoder in the stereo scenario, hence we are only interested in the relative performance of the fused SIs against the standard Extrapolation SI.

Secondly, for practical reasons, the second view is not WZ coded but intra coded, in other words all the frames used for producing Y_{mvS} and Y_{dp} are intra coded, in order to make a fair comparison and give some insights on the behavior on a scenario similar to the one in Figure 2, also the extrapolation is produced using intra coded frames.

The quantization matrices employed are the ones used in the DISCOVER project[13], the key frame coding is also the one used for the *Foreman* sequence in the DISCOVER Project. The RD points examined are Q1, Q4, Q7, Q8.

The sequences used are the stereo sequences provided by Microsoft Research[14], which are sequences having resolution 320x240 pixels, at 15 frames per second. The chosen sequences are IU, AC, IUJW and VK. We report the results only for luminance and only for the WZ frames for the right views of the 4 chosen sequences.

We also present the results in the form of Bjøntegaard[15] PSNR difference between the extrapolation curve and the other in the following table. It can be noted that the system using all the analyzed SIs is always better than the other two, verifying that the Multi-hypothesis decoder is robust in this scenario when the number of SIs grows.

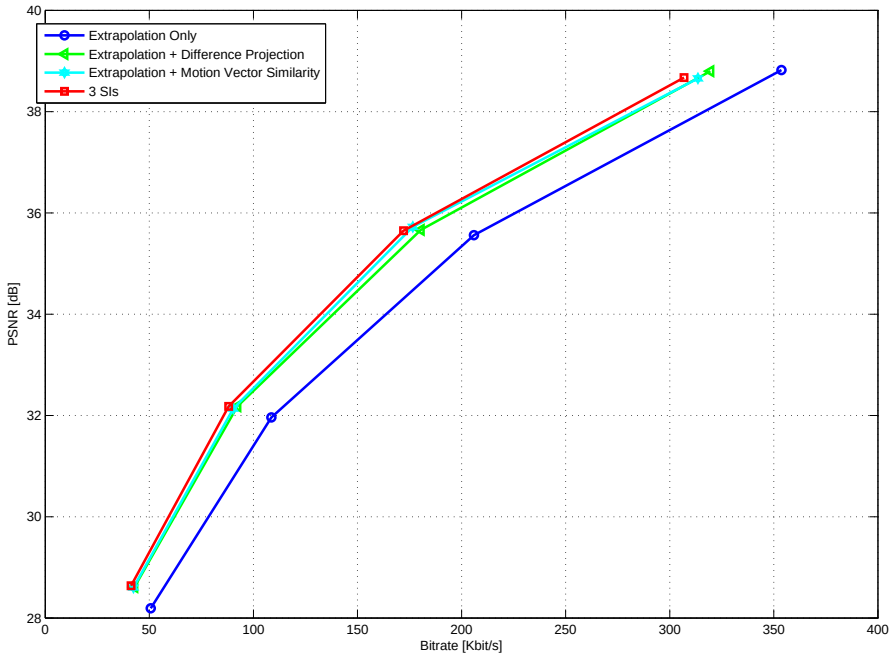

Figure 7. RD curves for sequence IU, WZ frames only, 15 fps, right view.

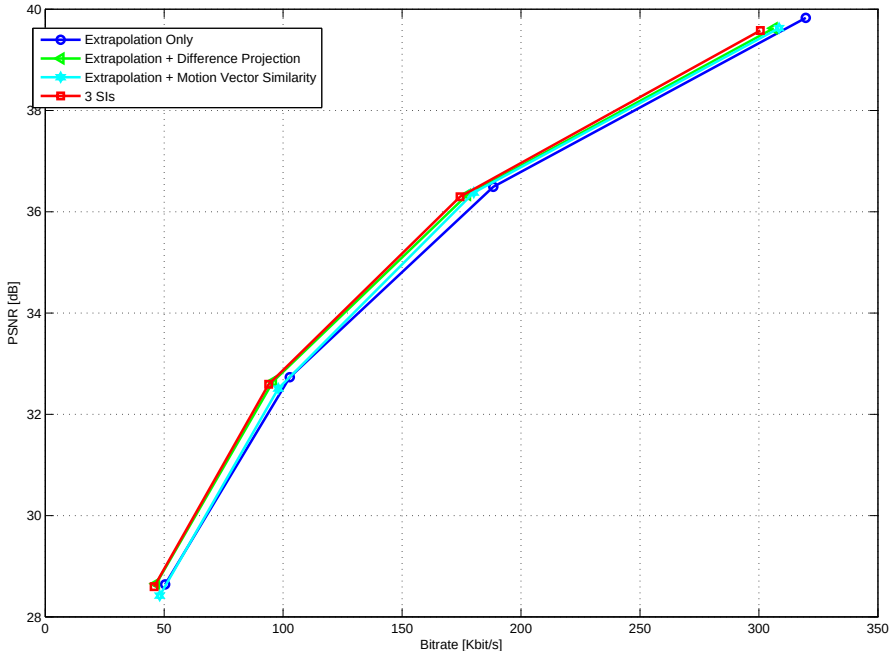

Figure 8. RD curves for sequence AC, WZ frames only, 15 fps, right view.

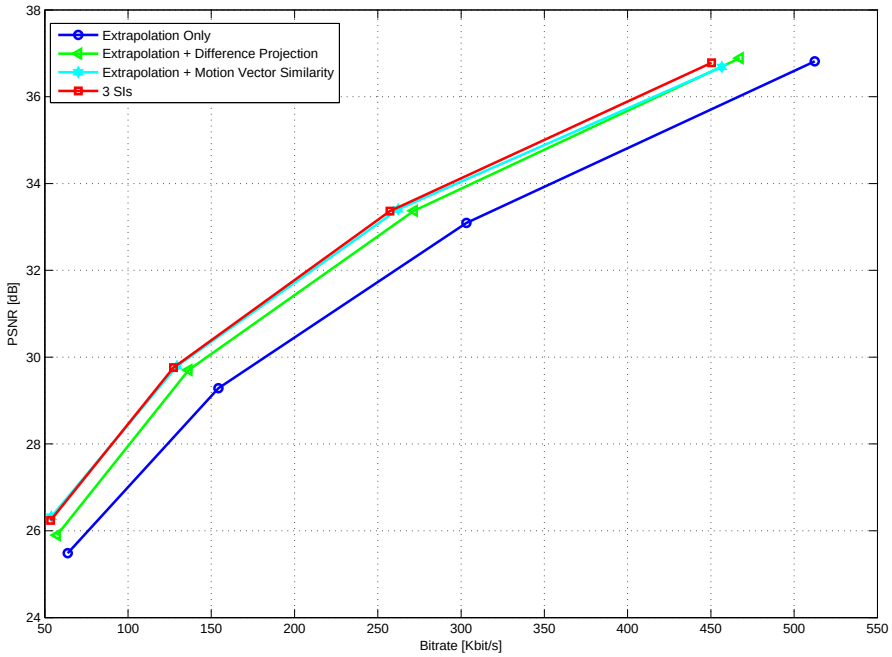

Figure 9. RD curves for sequence IUJW, WZ frames only, 15 fps, right view.

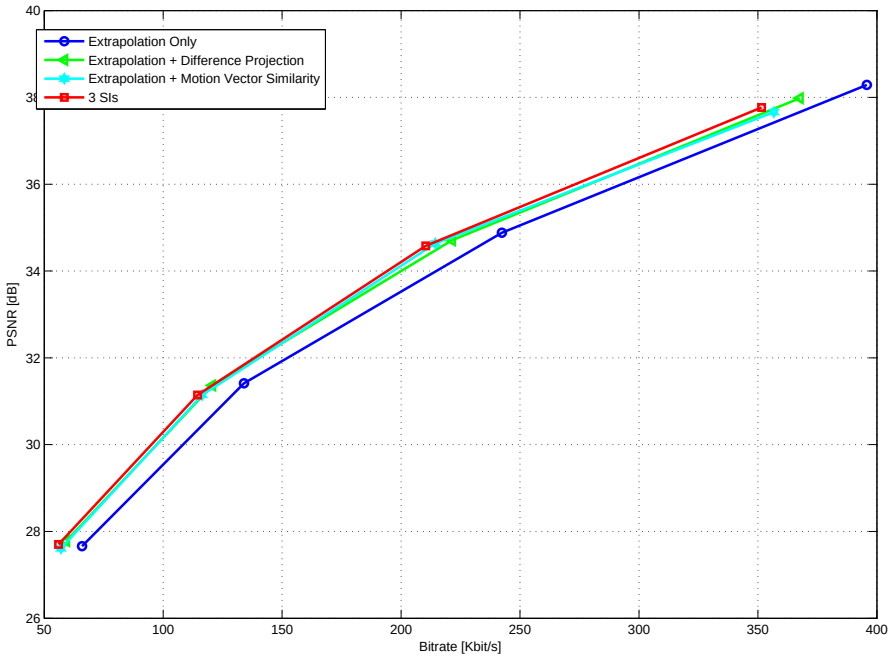

Figure 10. RD curves for sequence VK, WZ frames only, 15 fps, right view.

	IU	AC	IUJW	VK	Mean
Difference Projection	0.986	0.308	0.912	0.490	0.674
MV Similarity	1.033	0.073	1.240	0.479	0.706
3 SI	1.171	0.357	1.290	0.609	0.857

Table 1. PSNR[dB] Bjøntegaard Distance

5. CONCLUSIONS

In this paper applying DVC to the Stereo Low-delay scenario is addressed. Two techniques have been studied in order to obtain good quality SI. The SIs produced have been fused together using a Multi-Hypothesis decoder in order to have a system, which is able to automatically adapt to the changes in motion of the video sequence. It has been demonstrated that even with the low quality of the SI in the low-delay scenario, good results can be achieved compared to the extrapolation-based system. The two proposed SI generation systems perform in a different manner: while the difference projection has lower average performance it performs acceptably in all the sequences, on the other hand the MV similarity based method performs better on average but on the AC sequence the improvement is negligible. The multi-hypothesis decoder allowed the system to combine the best aspects of both the SIs in an automatic manner. In the multiview scenario the use of a multi-hypothesis decoder could solve the critical problem of how the fusion between the various SIs should be performed.

In the future we will concentrate on developing a real residual estimation module for these techniques and we will explore the use of multi-hypothesis decoders also in other multiview scenario with both extrapolation and interpolation based methods.

ACKNOWLEDGMENTS

The authors would like to thank David Varodayan[10] for helping us with the production of the LDPCA code having block length 4800.

REFERENCES

[1] Girod, B., Aaron, A. M., Rane, S., and Rebollo-Monedero, D., "Distributed Video Coding," *Proceedings of the IEEE* **93**, 71–83 (Jan. 2005).

[2] Puri, R., Majumdar, A., and Ramchandran, K., "PRISM: A Video Coding Paradigm With Motion Estimation at the Decoder," *IEEE Transactions on Image Processing* **16**, 2436–2448 (Oct. 2007).

[3] Slepian, D. and Wolf, J., "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory* **19**, 471 – 480 (jul 1973).

[4] Wyner, A. D. and Ziv, J., "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory* **22**, 1–10 (1976).

[5] Ouaret, M., Dufaux, F., and Ebrahimi, T., "Iterative multiview side information for enhanced reconstruction in distributed video coding," *J. Image Video Process.* **2009**, 3:1–3:17 (Jan. 2009).

[6] Huang, X., Brites, C., Ascenso, J. a., Pereira, F., and Forchhammer, S., "Distributed video coding with multiple side information," in [*Proceedings of the 27th conference on Picture Coding Symposium*], *PCS 2009*, 385–388, IEEE Press, Piscataway, NJ, USA (2009).

[7] Chen, Y., Cai, C., and Ma, K.-K., "Stereoscopic video error concealment for missing frame recovery using disparity-based frame difference projection," in [*Image Processing (ICIP), 2009 16th IEEE International Conference on*], *ICIP 2009*, 4289 –4292 (nov. 2009).

[8] Areia, J., Ascenso, J., Brites, C., and Pereira, F., "Wyner-ziv stereo video coding using a side information fusion approach," in [*Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*], *MMSP 2007*, 453 –456 (oct. 2007).

[9] Huang, X. and Forchhammer, S., "Cross-band noise model refinement for transform domain wynerziv video coding," *Signal Processing: Image Communication* **27**(1), 16 – 30 (2012).

[10] Varodayan, D., Aaron, A., and Girod, B., "Rate-adaptive codes for distributed source coding," *EURASIP Signal Processing Journal* **86**, 3123–3130 (Nov. 2006).

[11] Yang, S., Zhao, Y., Wang, S., and Chen, H., "Error concealment for stereoscopic video using illumination compensation," *Consumer Electronics, IEEE Trans.* **57**, 1907 –1914 (november 2011).

[12] Huang, X., Raket, L., Luong, H. V., Nielsen, M., Lauze, F., and Forchhammer, S., "Multi-hypothesis transform domain wyner-ziv video coding including optical flow," in [*Multimedia Signal Processing (MMSP), 2011 IEEE 13th International Workshop on*], *MMSP 2011*, 1 –6 (oct. 2011).

[13] "Discover project test conditions," (December 2007). http://www.img.lx.it.pt/ discover/test_conditions.html.

[14] "Microsoft research stereo video database." http://research.microsoft.com/en-us/projects/i2i/data.aspx.

[15] Bjøntegaard, G., "Calculation of average psnr differences between rd curves," *ITU-T Q6/SG16, Doc. VCEG-M33, in: 13th Meeting, Austin, USA, April* (2001).