Mixed resolution framework for distributed multiview coding

Diogo C. Garcia[a], Camilo C. Dórea[a], Bruno Macchiavello[a], Ricardo de Queiroz[a] and Debargha Mukherjee[b]

[a]Departamento de Engenharia Elétrica, Universidade de Brasília, 70.910-900, Brasília, DF, Brazil;
[b]HP Laboratories, Media Technologies Laboratory, CA 94304, Palo Alto, USA

ABSTRACT

This work presents a new distributed multiview coding framework, based on the H.264/AVC standard operating with mixed resolution frames. It allows for a scalable complexity transfer from the encoder to the decoder, which is particularly suited for low-power video applications, such as multiview surveillance systems. Greater quality sequences are generated by exploiting the spatial and temporal correlation between views at the decoder. The results show a good potential for objective quality improvement over simulcast coding, with no extra rate cost.

Keywords: Distributed video coding, multiview, side information.

1. INTRODUCTION

In recent years, much effort has been dedicated to the development of distributed video coding (DVC) techniques.[1] The main goal is to reduce the encoder complexity, transferring to the decoder the task of finding the correlation between frames. In conventional video coding techniques, such as the H.264/AVC standard,[2] motion estimation and compensation at the encoder offer some of highest gains in compression. On the other hand, DVC frameworks tend to encode frames independently, to reduce the encoder complexity. Using joint decoding of frames, these frameworks increase the coding gain to levels similar to those of conventional coding. These results are based on the theorems of Slepian-Wolf[3] and Wyner-Ziv.[4]

In settings where multiple cameras are used to register the same scene, the conventional coding approach, multiview coding (MVC), searches for temporal correlations (between frames of the same camera) as well as interview correlations (between frames of different cameras). Significant gains are obtained in this fashion, as opposed to coding all views independently from each other.[5] However, it also implies a higher complexity burden for the encoder. Furthermore, MVC imposes on the encoder the access to all video sources, which may be inviable to some applications.

Distributed multiview coding (DMC) is capable of reducing the work load on encoders, being suitable in scenarios of multiple low-power cameras. Nevertheless, it is also of interest for the decoder to have a choice of operating not only at high complexity mode, but also at low complexity.

The most common DVC frameworks use key frames coded in Intra mode,[2] and Wyner-Ziv (WZ) frames in between.[1] Without motion compensation and estimation, coding is substantially less complex. The decoder uses the key frames, and possibly motion estimation (and interview analysis for the DMC case), to create an estimate of the WZ frames (the so-called side information).

DMC techniques[6–9] transfer to the decoder the tasks of interview disparity estimation, view rectification and interpolation, occlusion handling, and others. In order to increase or decrease decoder complexity, these techniques can only vary the proportion between key and WZ frames, i.e., temporal scalability. A low complexity decoding must rely only on the key frames.

A mixed resolution DVC framework[10] is applied in this work for DMC. It allows scalable complexity transfer to the decoder, using lower resolution (decimated) frames, as well as key and WZ frames. A typical application for this DVC framework is real-time video transmission by cellular phones and security cameras, where a complexity increase results in higher battery power consumption. If the end user is another cellular phone, the decoding

E-mail: {diogo,camilo,bruno}@image.unb.br, queiroz@ieee.org, debargha.mukherjee@hp.com

can also be made with low complexity. If there are more available resources for decoding, such as in personal computer or a server, the framework allows for quality improvement for the decoded sequence.

Most of the previously referenced DMC techniques[6,7,9] choose one view as base for the coding of the other views. This assymmetry results in unequal resource allocation in coding, and also in differences in perceptual and objective quality between views. The proposed DMC framework is symmetric in this sense.

This paper is organized as follows. The proposed MR-DMC framework is described in Section 2. Test conditions and experimental results are shown in Section 3. Conclusions and future lines of research are discussed in Section 4.

2. FRAMEWORK

2.1 MR-DVC Framework

The proposed DMC framework is based on the single view mixed resolution DVC framework,[10] which presents a video codec with scalable complexity, implemented on the H.264/AVC codec. It uses mixed resolution, that is, a combination of frames with normal and reduced resolution. As shown in Figure 1, spatially reduced (decimated) frames are inserted between normal resolution key frames, such that motion estimation and compensation is less complex in the first kind of frame. In this way, the encoder complexity is diminished in proportion to the chosen degree of decimation and the number of decimated frames, providing greater control over the transfer of complexity from the encoder to the decoder. Finally, an enhancement layer is coded with Wyner-Ziv techniques[1,10] for the reduced frames, denominated non-reference Wyner-Ziv (NRWZ) frames.

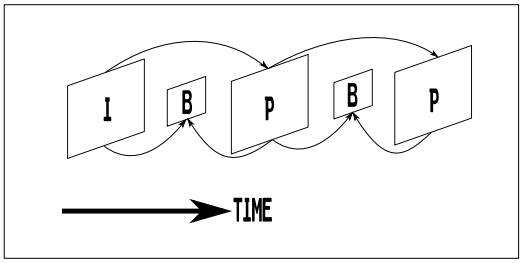


Figure 1. MR-DVC framework with one key frame (normal resolution) for each NRWZ frame (reduced resolution).

The compression is based on the H.264/AVC standard,[2] which defines three main kinds of frames: intra (I - spatial prediction), inter (P - spatial and temporal prediction, based on a previous frame) and bipredictive (B - spatial and temporal prediction, based on previous and later frames). In the MR-DVC framework, key frames can be of the I and P kinds, and NRWZ frames can be of the P and B kinds.

The decoder has the choice of decoding key and NRWZ frames, and ignore the enhancement layer, or generate better quality versions of NRWZ frames using key frames and the enhacement layer. In the former case, the decoding has low complexity, and in the latter, it has high complexity. In this manner, the codec presents scalable complexity, since the encoder can choose to use NRWZ frames, and the decoder can choose to enhance these frames or not.

The enhancement layer corresponds to the Wyner-Ziv coding of the residue of the NRWZ frames. The encoder calculates the difference between the original frame and the NRWZ frame interpolated to the original dimensions, takes the DCT transform of this residue, quantizes the output coefficients and calculates the corresponding memoryless cosets.[10]

At the decoder, the enhancement layer can only be used if there is some approximation of the original frame, since the cosets indicate the difference between this and the NRWZ frames. In order to do so, a process of semi super resolution is used, consisting of adding high-frequency components from the key frames to the NRWZ frames, approximating the original frame (denominated side information). This process is called side information generation.

The reference frames for the semi super resolution are: the NRWZ frame interpolated to the original sizes of the frame (LR-NRWZ frame); and the key frames decimated and interpolated back to the original dimensions

(LR-K frames), which represent low-pass versions of the key frames. The high frequency of these frames is the residue between them and the LR-K frames. To add high frequency information to NRWZ frames, a process of motion estimation is made for the LR-NRWZ frames, using LR-K as reference. The corresponding motion vectors point to the positions in the key frames where the high frequency information should be obtained and added to the NRWZ frames.

2.2 MR-DMC Framework

As in MR-DVC, the MR-DMC framework compresses multiview sequences transferring complexity from the encoder to the decoder. The encoding is simulcast, that is, each view is encoded separately, independent from each other. At the decoder, the side information is generated using key frames from the same view and from the other cameras as well. In this manner, the correlation between views can be explored to improve the semi super resolution process, in addition to the temporal correlation used by simulcast decoding.

In Figures 2 and 3, the arrows indicate which key frames are used as reference in the side information generation. Figure 2 presents the MR-DMC framework applied to stereo sequences (two simultaneous views), with a ratio of 1:1 for key and NRWZ frames. Figure 3 presents a 1:2 ratio in the same context. This ratio will here be denominated Wyner-Ziv ratio (WZR). Observe that the period of the key frames has been delayed in one frame for the right side camera, so that in the same instant, there are no two key frames. This allows for a better use of the correlation among views, because the reference key frames are closer in space and time to the NRWZ frames.

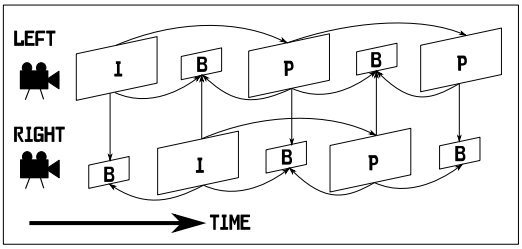


Figure 2. Stereo sequence decoded with a 1:1 WZR

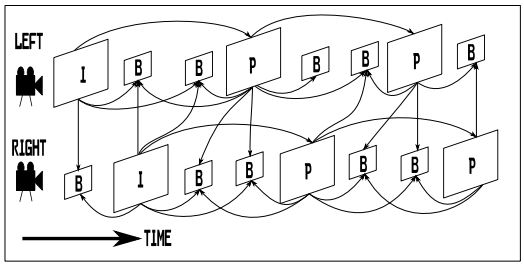


Figure 3. Stereo sequence decoded with a 1:2 WZR

Multiview codec applications improve the prediction between views based on the geometrical relationship among the cameras. Some of the main contributions are luminance compensation from different views, view compensation and occlusion handling.[11,12] Initially, our proposed framework takes advantage of the temporal and spatial correlations between 8x8 blocks within each of the views. At a later stage of development, it will handle dense disparity and/or geometry between views.

3. EXPERIMENTAL RESULTS

The side information generation of the proposed framework was tested with three multiview sequences publicly available,[13] '*ballroom*', '*exit*' and '*vassar*'. For each sequence, the first two views and all 250 frames were considered. Two Wyner-Ziv ratios were used, 1:1 and 1:2, with key frames in the IPPP mode and NRWZ frames in the B mode. The quantization parameters used were 22, 27, 32 and 37. A decimation factor of two was used

in the NRWZ frames. The same sequences were decoded in simulcast mode, for comparison. The Wyner-Ziv enhancement layer was not added in these initial experiments.

Table 1 presents the mean overall PSNR gains (luminance and chroma) in the side information generation for multiview decoding, as opposed to the simulcast decoding, for a WZR of 1:1 and the chosen quantization parameters. The results for a 1:2 WZR are presented in Table 2. It can be seen that for higher quantization parameters the smaller the gains become, but they are always positive (that is, there is no loss in using the multiview decoding). Furthermore, the multiview decoding presents very similar gains both for 1:1 and 1:2 WZRs.

Table 1. Mean PSNR gains in side information for multiview decoding, in relation to the simulcast decoding, for a 1:1 WZR

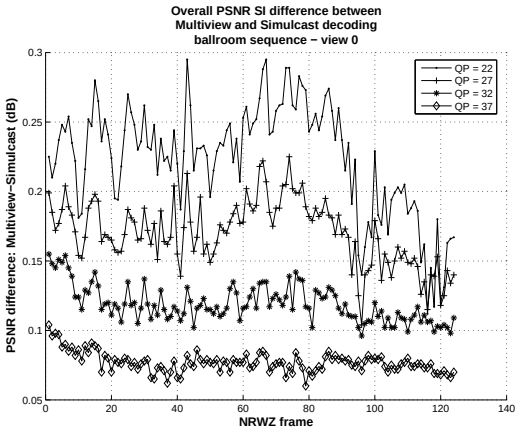
Sequence	QP = 22	QP = 27	QP = 32	QP = 37
ballroom - view 0	0.22 dB	0.17 dB	0.12 dB	0.08 dB
ballroom - view 1	0.16 dB	0.14 dB	0.10 dB	0.07 dB
exit - view 0	0.14 dB	0.10 dB	0.09 dB	0.07 dB
exit - view 1	0.12 dB	0.09 dB	0.07 dB	0.05 dB
vassar - view 0	0.16 dB	0.13 dB	0.10 dB	0.06 dB
vassar - view 1	0.12 dB	0.11 dB	0.09 dB	0.06 dB

Table 2. Mean PSNR gains in side information for multiview decoding, in relation to the simulcast decoding, for a 1:2 WZR

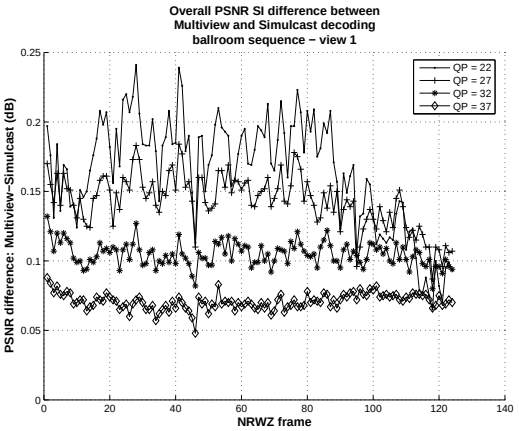
Sequence	QP = 22	QP = 27	QP = 32	QP = 37
ballroom - view 0	0.23 dB	0.17 dB	0.12 dB	0.08 dB
ballroom - view 1	0.16 dB	0.14 dB	0.10 dB	0.07 dB
exit - view 0	0.13 dB	0.11 dB	0.09 dB	0.07 dB
exit - view 1	0.13 dB	0.09 dB	0.07 dB	0.05 dB
vassar - view 0	0.16 dB	0.13 dB	0.10 dB	0.06 dB
vassar - view 1	0.12 dB	0.11 dB	0.09 dB	0.06 dB

Figure 4 shows the behavior of the afore-mentioned side information PSNR gain for each NRWZ frame, for all sequences under a 1:1 WZR. Figure 5 shows the same behavior for a WZR of 1:2. The multiview decoding has a higher side information PSNR than the simulcast decoding in all frames, for all quantization parameters. The observed gains are expected to increase with the inclusion of inter-view compensation techniques and dense subpixel disparity estimation.

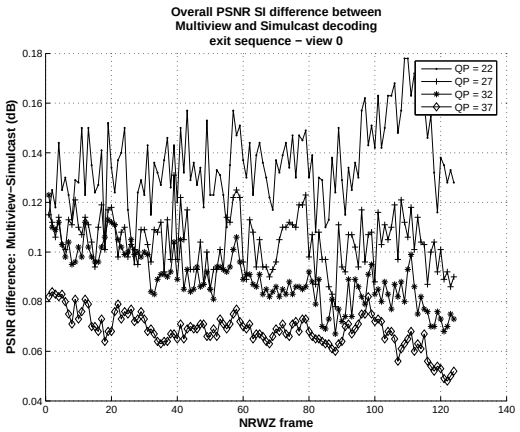
Figures 6 shows the rate-distortion curves for the side information generated in simulcast and multiview decoding, for all sequences and Wyner-Ziv ratios. It can be seen that the 1:1 WZR outperforms the 1:2 WZR both in simulcast and multiview decoding, which is the expected trade-off for lowering the encoder complexity while adding NRWZ frames. In both cases, there is a large margin for gain in side information generation to be explored, with the aforementioned multiview techniques.



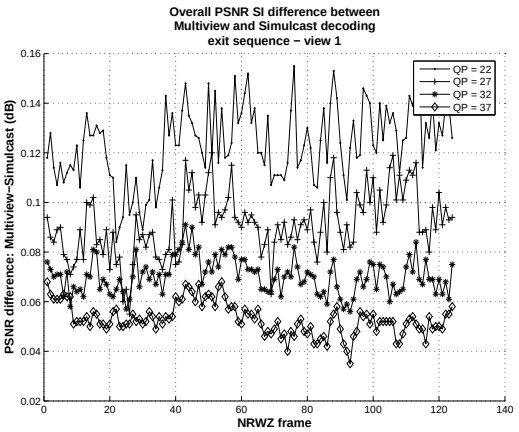
(a) Ballroom - View 0



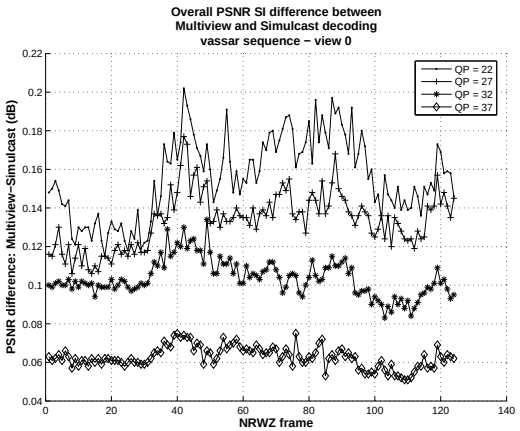
(b) Ballroom - View 1



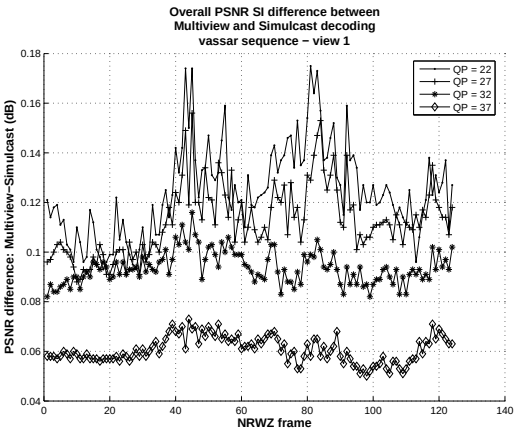
(c) Exit - View 0



(d) Exit - View 1

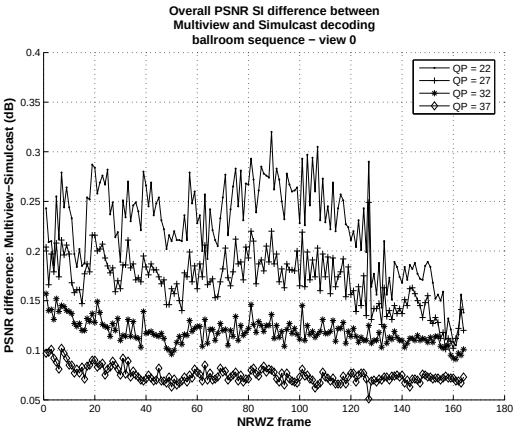


(e) Vassar - View 0

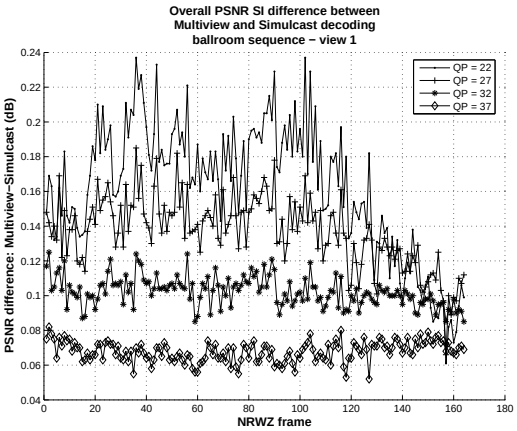


(f) Vassar - View 1

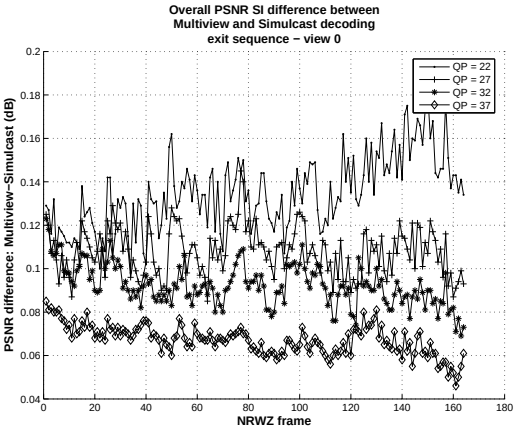
Figure 4. PSNR gains in side information for multiview decoding, in relation to the simulcast decoding, with a 1:1 WZR



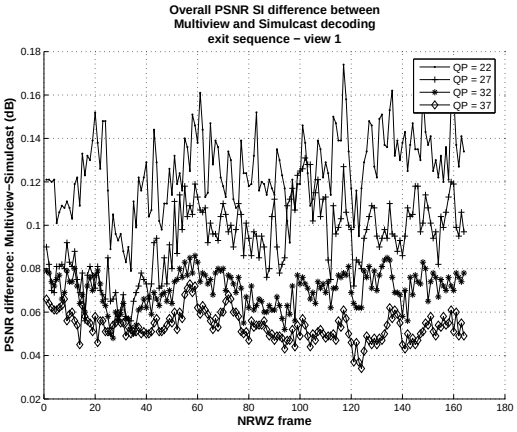
(a) Ballroom - View 0



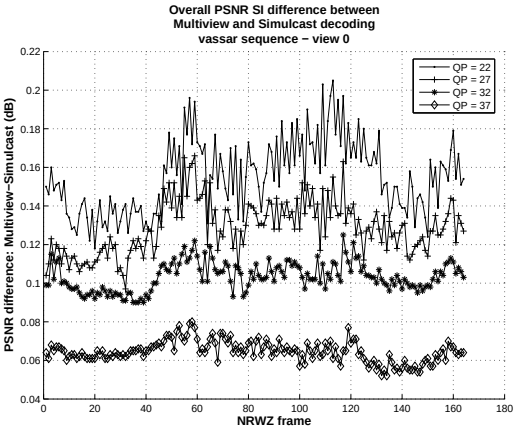
(b) Ballroom - View 1



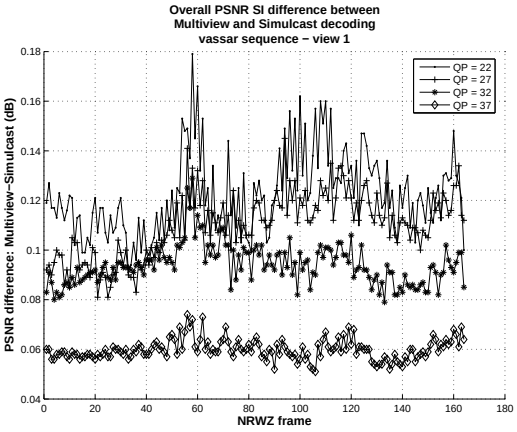
(c) Exit - View 0



(d) Exit - View 1

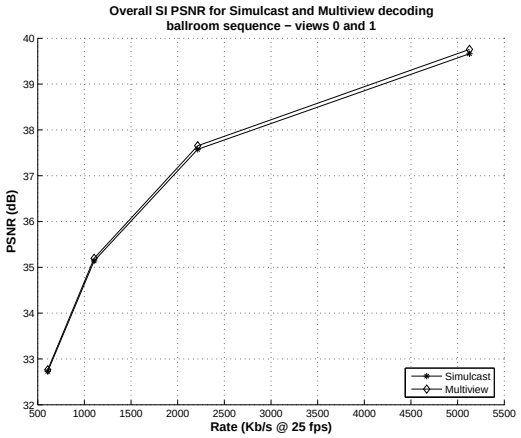


(e) Vassar - View 0

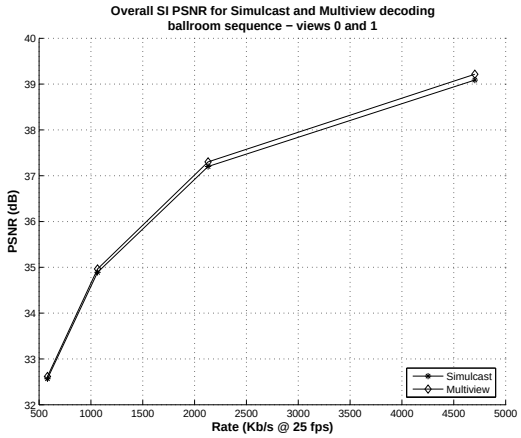


(f) Vassar - View 1

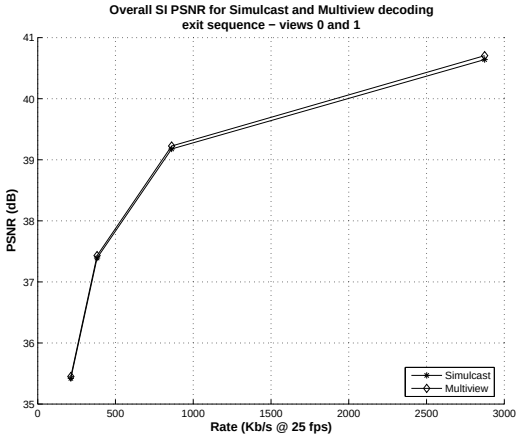
Figure 5. PSNR gains in side information for multiview decoding, in relation to the simulcast decoding, with a 1:2 WZR



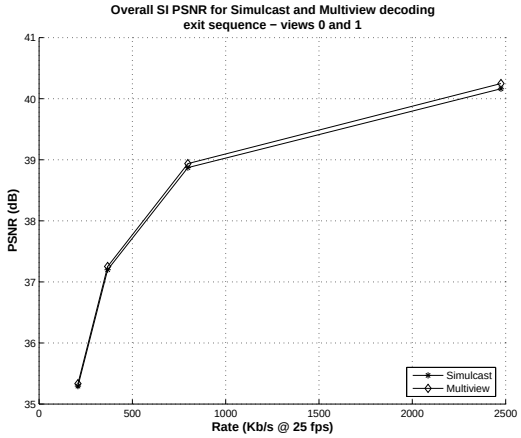
(a) Ballroom - 1:1 WZR



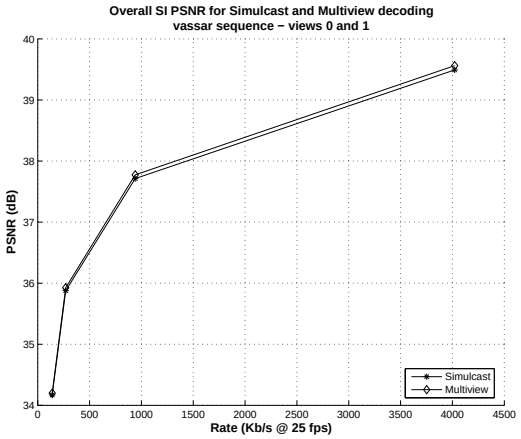
(b) Ballroom - 1:2 WZR



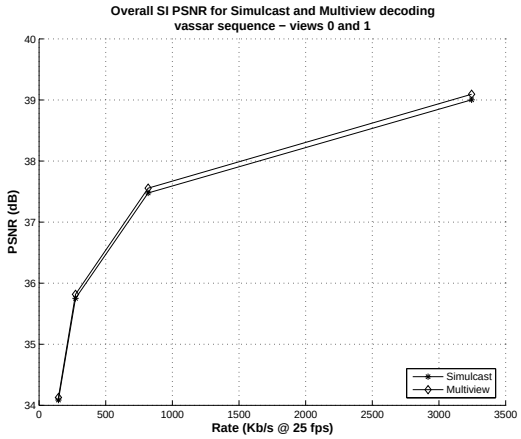
(c) Exit - 1:1 WZR



(d) Exit - 1:2 WZR



(e) Vassar - 1:1 WZR



(f) Vassar - 1:2 WZR

Figure 6. Rate-distortion curves for the side information with simulcast and multiview decoding

4. CONCLUSIONS

This paper presented a new framework for distributed multiview coding, based on the H.264/AVC standard and mixed resolution frames. The framework offers greater complexity scalability in coding and decoding, has symmetry in resource use for the encoder, and exploits spatial and temporal interview correlations in the decoding. The results show an objective quality gain for low complexity multiview encoders.

Higher gains are expected by testing the framework with more than two views, so that the correlation between views is better explored. The test sequences, for instance, offer eight views of the same scene. Next, the framework will be modified to take the geometrical relationship between cameras into acount. Techniques such as view correction, view interpolation and occlusion handling should contribute to a better side information generation.

REFERENCES

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

[2] Wiegand, T., Sullivan, G. J., Bjontegaard, G., and Luthra, A., "Overview of the h.264/avc video coding standard," *Circuits and Systems for Video Technology, IEEE Transactions on* **13**(7), 560–576 (2003).

[3] Slepian, D. and Wolf, J., "Noiseless coding of correlated information sources," *Information Theory, IEEE Transactions on* **19**(4), 471–480 (1973).

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

[5] Merkle, P., Smolic, A., Muller, K., and Wiegand, T., "Efficient prediction structures for multiview video coding," *Circuits and Systems for Video Technology, IEEE Transactions on* **17**, 1461–1473 (Nov. 2007).

[6] Artigas, X., Angeli, E., and Torres, L., "Side information generation for multiview distributed video coding using a fusion approach," in [*Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic*], 250–253 (June 2006).

[7] Ouaret, M., Dufaux, F., and Ebrahimi, T., "Fusion-based multiview distributed video coding," in [*VSSN '06: Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*], 139–144, ACM, New York, NY, USA (2006).

[8] Tagliasacchi, M., Prandi, G., and Tubaro, S., "Symmetric distributed coding of stereo video sequences," in [*Image Processing, 2007. ICIP 2007. IEEE International Conference on*], **2**, II –29–II –32 (16 2007-Oct. 19 2007).

[9] Guo, X., Lu, Y., Wu, F., Zhao, D., and Gao, W., "Wyner-Ziv-based multiview video coding," *Circuits and Systems for Video Technology, IEEE Transactions on* **18**, 713–724 (June 2008).

[10] Macchiavello, B., Mukherjee, D., and Queiroz, R. L., "Iterative side-information generation in a mixed resolution Wyner-Ziv framework.," *To be published in IEEE Trans on Circuits and Systems for Video Technology* (2009).

[11] Hur, J.-H., Cho, S., and Lee, Y.-L., "Adaptive local illumination change compensation method for H.264/AVC-based multiview video coding," *Circuits and Systems for Video Technology, IEEE Transactions on* **17**, 1496–1505 (Nov. 2007).

[12] Yamamoto, K., Kitahara, M., Kimata, H., Yendo, T., Fujii, T., Tanimoto, M., Shimizu, S., Kamikura, K., and Yashima, Y., "Multiview video coding using view interpolation and color correction," *Circuits and Systems for Video Technology, IEEE Transactions on* **17**, 1436–1449 (Nov. 2007).

[13] MERL, "Mitsubishi Electric Research Laboratories multiview video sequences." ftp://ftp.merl.com/pub/avetro/mvc-testseq.