VIEW SYNTHESIS MOTION ESTIMATION FOR MULTIVIEW DISTRIBUTED VIDEO CODING

Shinya Shimizu, and Hideaki Kimata

NTT Cyber Space Laboratories, Nippon Telegraph and Telephone Corporation
1-1 Hikari-no-oka, Yokosuka, Kanagawa, 239-0847, JAPAN
phone: + (81) 46-859-2703, fax: + (81) 46-859-2829, email: shimizu.shinya@lab.ntt.co.jp

ABSTRACT

Distributed Video Coding (DVC) is an attractive video coding scheme. It is well-known that the quality of the side information (SI) strongly impacts the coding performance of DVC. One of the most popular SI generation techniques is motion compensated temporal interpolation, where temporal interpolation is performed by assuming linear uniform motion. However, it is difficult to generate high quality SI because there are so many irregular motions. In multiview DVC, it becomes possible to utilize inter-view correlation in addition to temporal correlation. Therefore, this paper proposes a temporal frame interpolation method that can compensate irregular motion by estimating motion on view interpolated frames. Simulations show that the proposed method improves SI quality by up to 4.5 dB.

1. INTRODUCTION

Multiview video is attracting a lot of interest for many advanced visual media applications such as Free-viewpoint Television (FTV) and 3D Video [1, 2], and video surveillance. The recent remarkable advances in multiview video processing make it possible to realize such applications in the near future. The reduced cost of cameras is another factor stimulating multiview video systems.

FTV allows viewers to roam the captured scenes without concern for the positions of the real cameras; current video systems show the viewer just the view of the selected camera. One promising FTV application is video surveillance. View point freedom is desirable to increase the accuracy of object and event detection. One approach to FTV implementation is using multiview video and virtual view synthesis techniques.

In spite of the recent advances in related technologies, the number of views required for these applications is still large. Since the data amount is basically proportional to the number of views, achieving efficient compression is one of the most important issues in such applications.

Recently, the Joint Video Team (JVT) of ISO/IEC JTC1/SC29/WG11 Moving Picture Experts Group (MPEG) and ITU-T SG16 WP3 Q.6 Video Coding Experts Group (VCEG) released an amendment of MPEG-4 AVC/H.264 for multiview video coding (MVC) [3]. MVC can realize the efficient compression of multiview video by exploiting inter-view correlations. However, the computational complexity of the MVC encoder is extremely high because it must process multiple videos at the same time. It may be possible to overcome this problem by assigning one processor to each view. However this kind of implementation brings another problem; it is necessary to transmit local decoded pictures among processors. This may introduce additional delays and networking bottlenecks, especially for distributed camera networks.

Multiview distributed video coding (MDVC) is a solution that can achieve efficient compression, low complexity encoding, and no communication among views at the same time [4, 5]. In MDVC, the inter-view correlations are exploited only at the decoder by generating estimates, called side information (SI), in various ways. It has been proven that the coding performance strongly depends on SI quality. Therefore, improving SI quality is a major research topic. View interpolation (VI), sometimes referred to as view synthesis (VS), is a promising interview prediction method [6]. Although VI has the ability to compensate scene disparities precisely by using intrinsic and extrinsic camera parameters of views, it has been reported that the quality of VI side information (VISI) is very limited [7].

In MDVC, it is possible to use temporal SI, which is used in general DVC. Temporal SI is generated by assuming the linear motion model and inter-view SI assumes the Lambert reflection of objects. However it is difficult to interpolate frames because motions have the strong irregularities and no object has a true Lambert surface. Therefore, many fusion techniques linking intra-view and inter-view SI have been proposed that attempt to compensate these drawbacks [6, 8, 9]. However, the existing methods fail to improve the SI quality given non-Lambert objects with has irregular motion because these methods simply propose to average some SIs depending on the magnitude of the estimated motions, which is one of the reliability measures on interpolation correctness.

In this paper, we propose a novel method to generate high quality SI by utilizing both temporal and inter-view correlations. This paper is organized as follows. First, we overview MDVC and describe the issue treated in this paper in Section 2. The proposed method is introduced in Section 3. Simulation results are presented in Section 4. Finally, Section 5 concludes this paper with some remarks.

2. MULTIVIEW DISTRIBUTED VIDEO CODING

MDVC is the multiview version of distributed video coding (DVC), which is based on the theoretical results reported by Slepian & Wolf [10] (for the lossless case) and Wyner & Ziv [11] (for the lossy case). These papers revealed that the coding performance that results from the separate encoding of two correlated sources can match the performance of joint encoding if they are jointly decoded by exploiting the statistical dependencies. In a practical DVC implementation [12, 13], the frames are classified into two groups: the key frames (KFs) and the Wyner-Ziv frames (WZFs). KFs and WZFs are encoded separately. Only the parity bits of channel

 ISSN 2076-1465

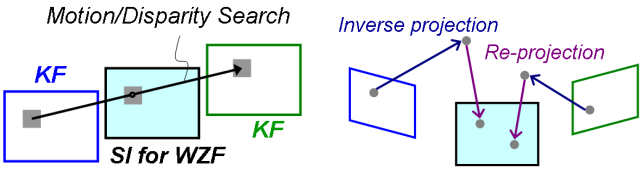

Figure 1: General Side Information Generation Method (left: MCTI/DCVP, right: VISI.

coded WZFs are transmitted for WZFs while source coding is applied to KFs. At the decoder, an estimate, called side information (SI), of WZF is generated by using the decoded KFs. The decoder considers the SI as WZFs with some channel errors, and these errors are corrected in channel decoding by using the received parity bits.

In the mono-view case, motion compensation temporal interpolation (MCTI) is used to generate the SI [14]. MCTI estimates the motions between KFs, and the resulting motions are interpolated by assuming linear uniform motion as shown Figure 1. In the multiview case, inter-view estimation is enabled in addition to the temporal one. Many methods have been proposed to utilize inter-view correlation.

One of the most popular inter-view SI generation methods is disparity compensation view prediction (DCVP) [8]. DCVP applies almost the same algorithm with MCTI to frames from different views. One modification is introduced: an optimal interpolation distance is computed at the first frame by using the decoded frame of all views, and then the weight obtained for interpolating disparities is used for the remaining frames. This modification is introduced because the view number gives incorrect information on view position while the temporal timestamp gives the exact temporal position. However, the disparities depend on not only the distance of the camera but also object position, so this scheme fails to compensate the disparities correctly. As a result, DCVP usually yields lower SI quality than MCTI.

View interpolation (synthesis) SI (VISI) can achieve precise disparity interpolation by simulating the camera shooting process; each pixel in KFs is inversely projected into the scene to reconstruct 3D points, and then the reconstructed points are re-projected into the camera plane of WZF as shown Figure 1. In order to perform the inverse projection, it is necessary to know the scene depth, the distance between camera and object. Therefore, depth estimation like stereo matching is conducted between KFs in the process of VISI generation. This method can interpolate scene geometry precisely if the depth information is correctly estimated.

It was reported that VISI quality is very limited [7]. There are many factors that prevent the existing VI methods from providing good predictions. One is the difficulty of estimating the geometric information correctly. This problem has been tackled for many years by a lot of researchers, especially in the computer vision field. The accuracy of estimation is not perfect and further studies are needed, but it seems acceptable for the purpose of estimating scenes as shown Figure 2. Therefore, this problem is not treated in this paper.

Another factor is the inability to compensate the inter-view image signal mismatch caused by the heterogeneous cameras. It is difficult to use identical camera settings in practice, so many kinds of interview image signal mismatch

Figure 2: Example of VISI (top left: original, top right: VISI, bottom: estimated depth map).

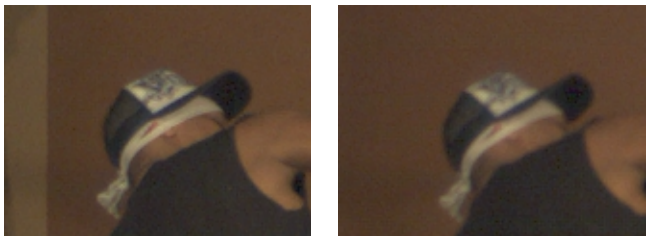

Figure 3: An example of inter-view focus mismatch.

occur. One example is the use of different exposure, focal length, and/or shutter speed settings, which can result in differences in in-focus position and range that appear as sharpness and blur disparities (see Figure 3). Another example is the use of different gain and dynamic range control settings, which result in image signals with different intensities. We have proposed adaptive filtered view interpolation SI (AFVISI) to reduce these inter-view signal mismatch [15]. Adaptive filtering can improve the SI quality when the quality of VISI is relatively high. However, if there noise is large, the filtering process sometimes spreads the noise into the neighbouring regions. Therefore, we propose another approach to improve SI quality through the use of VISI.

Multiview motion estimation (MVME) is generated by using both temporal KF and inter-view KF [16]. MVME estimates the motion vectors in the side views and then motion compensated prediction is performed at the center view. Before conducting motion compensation prediction, the estimated motion vectors are transformed by using the estimated disparity vector between side and center views. Figure 4 illustrates the MVME concept and one possible path for prediction. If there are 8 neighbouring KFs available, there are 8 paths to predict one WZF. Therefore, the final MVME is generated by averaging all 8 predicted images.

The proposed method also uses both temporal KFs and inter-view KFs. In this sense, the proposed method takes an approach similar to that of MVME. However the proposed method has no limitation in camera settings while MVME can provide correct vector transformation only when the optical axes of all cameras are orthogonal to the motion. More-

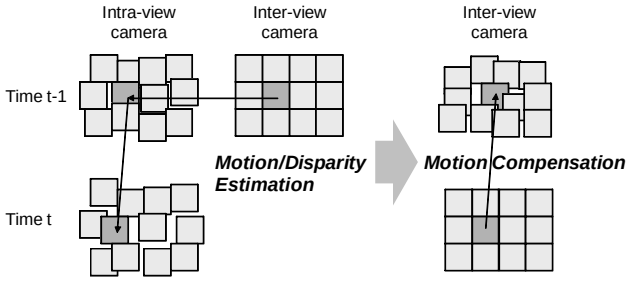

Figure 4: MVME scheme.

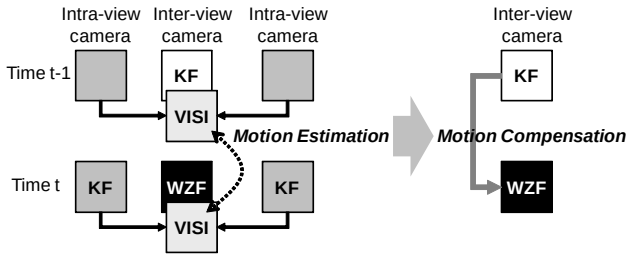

Figure 5: Basic concept for the VSME technique.

over, the proposed method uses only temporal image signals to generate the final SI while MVME uses both temporal and inter-view.

3. VIEW SYNTHESIS MOTION ESTIMATION

As already described, one of the fatal drawbacks of MCTI, which interpolates frames temporally, is the assumption of linear uniform motion. On the other hand, VISI, which is one of the geometrically precise inter-view frame interpolation methods, fails to generate high quality SI due to its in ability to compensate the inter-view signal mismatch caused by heterogeneous camera settings and non-Lambert reflection. Therefore, we propose a novel SI generation method by combining these two methods while eliminating their defects. In other words, the proposed method predicts image signals temporally with no assumption of motion model.

The main idea behind the proposed method, View Synthesis Motion Estimation (VSME), is depicted in Figure 5. The motion vectors are estimated between VISIs, which are generated on both KF and WZF. Because VISI can achieve geometrically precise frame interpolation, it is possible to consider VISI on WZF as a correct estimate of the scene of WZF. Therefore, it becomes possible to estimate motions between KF and WZF with no assumption of a motion model by estimating motions between VISI for KF and VISI for WZF. It is also possible to conduct motion estimation between KF and VISI for WZF. However it is difficult to find precise motion vectors because of the signal mismatch between KF and VISI; the motion estimation errors degrade the resulting SI quality. The drop in SI quality can be offset by the low computational complexity because the process of VISI generation requires a lot of computation. Therefore, one future work is to establish an accurate motion estimation algorithm with KF and VISI. Details of VSME are explained below.

First, VISIs are generated for KFs and WZF; KFs are the temporal reference frames of motion compensated prediction. It is possible to apply any algorithm to interpolate inter-view frames. The most important requirement of this view synthesis process is providing geometrically correct predictions. In this paper, depth maps are estimated for the view synthesis target frame and simple 3D warping is used to synthesize the view [17]. Note that depth estimation is conducted at the decoder side, not at the encoder side: this means that no depth information is encoded. Depth maps $\mathbf{d}$ are estimated by minimizing the following cost function E;

$$E(\mathbf{d}) = \sum_{p} \left(I_L(d_p, p) - I_R(d_p p)\right)^2 + \lambda \sum_{\{p,q\} \in N} |d_p - d_q| \quad (1)$$

, p represents a pixel in the depth map, N represents the set of adjacent pixels, d_p represents the depth value of pixel p, and $I_L(d_p, p)$ and $I_R(d_p, p)$ represent the pixel value of the corresponding pixels of center view's pixel p with depth d_p on left and right views, respectively. This minimization problem is solved by graph cut.

Second, all generated VISIs are low pass filtered to improve the reliability of the motion vectors because warping-based view interpolation introduces artificial noise, especially at high frequencies. Next, a block matching algorithm is used to estimate the motion of each block in KF by using VISI for KF and VISI for WZF. The parameters are the block size, the search window size, and the search range.

Third, the weighted vector median filter is applied to increase spatial coherence of estimated motion vectors [18].

Last, motion compensated frame prediction is performed by using the obtained motion vector field and KF. If both forward and backward KFs are available, motion vector fields are estimated separately and then the bidirection motion compensation of standard video coding is performed to obtain the final SI.

4. EXPERIMENTS

The breakdancers and ballet multiview test sequences were used in the simulations [19]. The spatial resolution is 256x192 and temporal resolutions are 15 fps for both sequences. Both sequences contain rapid and random motions with relatively large static background.

There are many possibilities for the positions of the KFs and WZFs. We chose the simplest setting as illustrated in Figure 6. The views are separated into two categories: intra-view and inter-view. Inter-view is decoded jointly with the other views while intra-view is decoded without any other views. Every second view was defined as intra-view and the others as inter-view. All the frames in intra-view are encoded by H.264/AVC Intra. However the proposed scheme doesn't care whether the intra-views employ the mono-view DVC scheme or not as long as the frames that have the same timestamps with KFs and WZF, which are reference and target frames for frame interpolation, are available. We also assumed that KFs are placed at every second position, and encoded by H.264/AVC Intra. Note that it is easy to consider the case where the KF interval is longer. In the experiments below, only view 4 was evaluated because it is easy to extend the simulation to the other inter-views.

In the experiments, we evaluated just the SI quality. Since many SIs have been proposed for MDVC, fusion techniques have also been investigated to generate better SI by

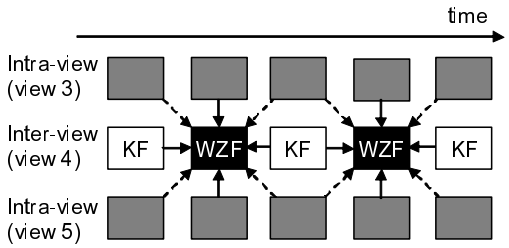

Figure 6: View and frame setting (Gray frames express Intra-view frames which can be either KFs of WZFs.

utilizing the advantages of each SI[6, 8, 9]. Therefore, it is possible to obtain a better SI by taking the proposed VSME as one of the candidates for these fusion techniques. Iterative SI generation is another technique to increase SI quality [7]. It may be also possible to enhance the quality by using VSME as a part of the initial SI in the IMSI generation process.

In order to confirm the effectiveness of the proposed method, we compared the PSNR value of each SI against the originals. For the comparison, we implemented MCTI, DCVP, MVME, VISI, and AFVISI. All the settings for motion estimation were identical for all SIs. The average PSNR values are shown in Figure 7. MVME-Motion is MVME with only 4 motion paths, and MVME-All is with all 8 paths.

As can be seen, the proposed method always shows the best performance. The gains ranged from 4.5dB to 0.5 dB, and depend on the sequence and KF quality. The proposed method brings significant gains with high quality KFs, but relatively small gains with low quality KFs, especially for the breakdancers sequence. One of the reasons for the low performance with low quality KFs is the limited quality for the reference frame of motion compensated prediction. The difficulty of depth estimation with low quality KFs and intra-view frames could be considered as another factor, but the influence may be limited because the degradation of VISI, which also requires depth estimation process, is not so large relative to that of VSME. Figure 8 shows the examples of VISI and VSME. The example shows that precise view interpolation is not necessary to improve VSME quality.

We also investigated the rate-distortion performance by using one of the basic DVC frameworks called DCT-domain Wyner-Ziv codec. The used codec is almost the same as the well-known DISCOVER codec [20]. One difference is the parameter for modelling error distribution between WZF and SI, which is modelled as a Laplacian distribution. In the experiment, the optimal Laplacian distribution parameter was calculated from WZF and SI while the discover codec estimates it from KFs and SI.

Figure 9 show the RD performance. For each rate point, KFs and intra-view frames were encoded to have almost the same quality with the corresponding WZFs. As can be seen, the proposed method outperforms the other methods over all bitrates. The improvements reached about 2.5 dB for ballet and about 1.5 dB for breakdancers at the middle bitrate.

5. CONCLUSION

We proposed a novel temporal SI generation method for multiview distributed video coding. The proposed method, view

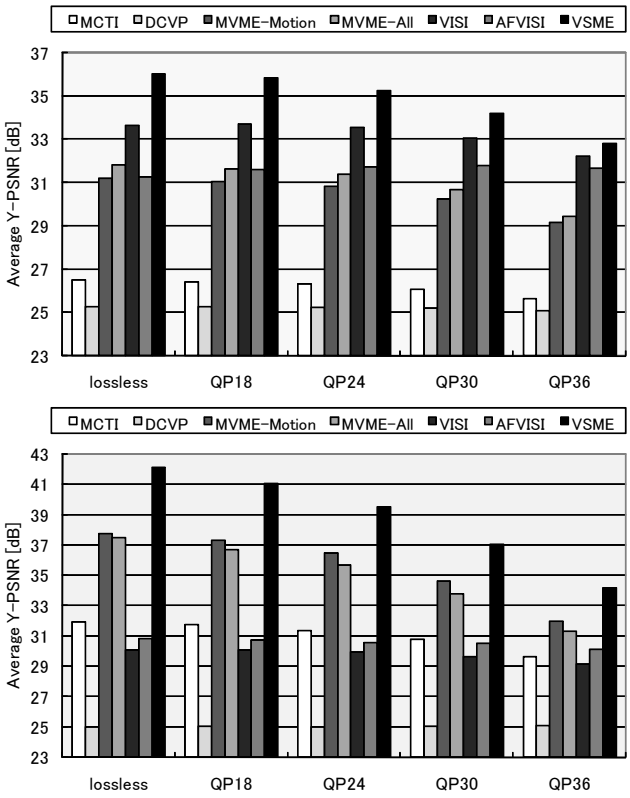

Figure 7: SI quarity(top:breakdancers, bottom:ballet).

synthesis motion estimation, temporally interpolates the image signals with no assumption of a motion model. In other words, the proposed method can compensate random motion correctly. In order to estimate such motion, view interpolated frames are generated for both KF and WZF. Experiments show that the proposed method improves SI quality by up to 4.5 dB.

One of the drawbacks of the proposed method is its computational complexity because the view interpolation process introduces extremely high computation loads. Therefore, one of the future works is to reduce the number of required view interpolations. One approach is to estimate motion with KF itself and view interpolated frame for WZF. It is also necessary to consider another algorithm that can generate high quality SI with lower quality KF.

REFERENCES

[1] M. Tanimoto, "Overview of free viewpoint television," *Signal Processing: Image Communication*, vol. 21, no. 6, pp. 454–461, July 2006.

[2] A. Smolic, K. Müller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand, "3d video and free viewpoint video - technologies, applications and mpeg standards," in *Proc. ICME2006*, July 2006, pp. 2161–2164.

[3] A. Vetro, P. Pandit, H. Kimata, A. Smolic, and Y.-K. Wang, "Joint draft 8.0 on multiview video coding," JVT Doc. JVT-AB204 (rev.1), July 2008.

[4] X. Zhu, A. Aaron, and B. Girod, "Distributed compression for large camera arrays," in *Proceedings of the*

Figure 8: Generated SIs (left: VISI, right: VSME).

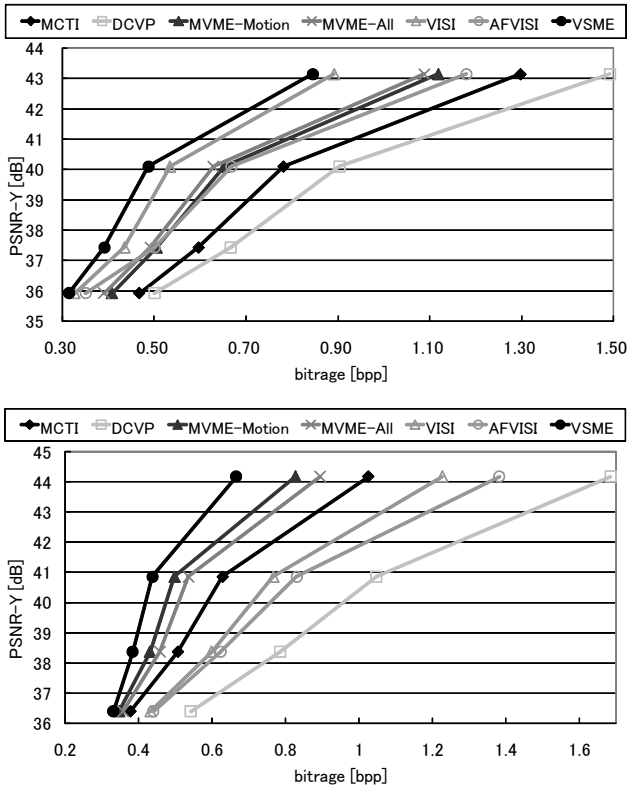

Figure 9: RD performance(top:breakdancers, bottom:ballet).

IEEE Workshop on Statistical Signal Processing, 2003, pp. 30–33.

[5] G. Toffetti, M. Tagliasacchi, M. Marcon, A. Sarti, S. Tubaro, and K. Ramchandran, "Image compression in a multi-camera system based on a distributed source coding approach," in *Processdings of European Signal Processing Conference 2005*, September 2005.

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

[7] M. Ouaret, F. Dufaux, and T. Ebrahimi, "Iterative multiview side information for enhanced reconstruction in distributed video coding," *EURASIP Journal on Image and Video Processing*, 2009.

[8] M. Ouaret, F. Dufaux, and T. Ebrahimi, "Multiview Distributed Video Coding with Encoder Driven Fusion," in *The 2007 European Signal Processing Conference (EUSIPCO-2007)*, 2007.

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

[10] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *Information Theory, IEEE Transactions on*, vol. 19, no. 4, pp. 471–480, July 1973.

[11] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *Information Theory, IEEE Transactions on*, vol. 22, no. 1, pp. 1–10, January 1976.

[12] R. Puri and K. Ramchandran, "Prism: A video coding architecture based on distributed compression principles," EECS Department, University of California, Berkeley, Tech. Rep. UCB/ERL M03/6, 2003.

[13] A. Aaron, R. Zhang, and B. Girod, "Wyner-ziv coding of motion video," in *Proc. Asilomar Conference on Signals and Systems*, 2002, pp. 240–244.

[14] J. Ascenso, C. Brites, and F. Pereira, "Improving frame interpolation with spatial motion smoothing for pixel domain distributed video coding," in *the 5th EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services*, July 2005.

[15] S. Shimizu, Y. Tonomura, H. Kimata, and Y. Ohtani, "Improved view interpolation for side information in multiview distributed video coding," in *Proceedings of Third ACM/IEEE International Conference on Distributed Smart Cameras*, 2009, pp. 1–8.

[16] X. Artigas, F. Tarres, and L. Torres, "A comparison of different side information generation methods for multiview distributed video coding," in *SIGMAP*, 2007, pp. 450–455.

[17] S. Yea and A. Vetro, "View synthesis prediction for rate-overhead reduction in ftv," in *Proc. 3DTV-Conference*, May 2008, pp. 145–148.

[18] L. Alparone, M. Barni, F. Bartolini, and V. Cappellini, "Adaptively weighted vector-median filters for motion-fields smoothing," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, pp. 2267 – 2270.

[19] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 600–608, 2004.

[20] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, "The discover codec: Architecture, techniques and evaluation," in *Proceedings of Picture Coding Symposium 2007*, 2007.