Stretching, Compression and Shearing Disparity Compensated Prediction techniques for Stereo and Multiview Video Coding

Ka-Man Wong, Lai-Man Po, Kwok-Wai Cheung[#], Ka-Ho Ng, and Xuyuan Xu

Department of Electronic Engineering, City University of Hong Kong, Hong Kong SAR, China
[#]Department of Computer Science, Chu Hai College of Higher Education, Hong Kong SAR, China

Abstract - In multiview video coding, disparity compensated prediction exploits the correlation among different views. A common approach is to use the conventional motion compensated prediction to predict disparity effect among different views. However, the same object in different views usually has deformation of different extents and, thus, accurate disparity prediction cannot be achieved with such simple translational motion model. Previous attempts to achieve more accurate disparity prediction are usually too complex for practical implementation. In this paper, stretching, compression and shearing (SCSH) effects are investigated to better model the disparity effect in disparity compensated prediction. To achieve SCSH effects with minimal computation, an efficient disparity compensated prediction using subsampled block-matching technique is proposed. No affine parameters estimation or additional frame buffers is required and the overall increase in memory requirement and computational complexity is moderate. Experimental results show that the new technique can achieve up to 4.84% bitrate reduction in inter-view prediction using JM17.0 reference software implementation.

Index Terms – Multiview Video Coding, Disparity Compensated Prediction, Stretching, Compression, Shearing.

I. Introduction

With the rapid development of image acquisition, video compression and video display technologies, 3D movies and 3DTV become feasible in recent years. The user's perception of depth and the associated sensation of reality provided by 3D videos are the most attractive features in digital entertainment. There are two popular types of 3D video -- stereo video and multiview video. Stereo video has two views, usually left and right, which emulate the stereoscopic vision of human to provide depth perception. Multiview video has two or more views with view angle chosen by user or automatic means. Various 3D display systems using different video display technologies are available to movie theaters and home entertainment market for 3D video display. Multiview video coding (MVC) is a key technology to enable efficient coding, storage and transmission of such video data [1].

Both MPEG-2 [2] and H.264/AVC [3] can support up to two views by interleaving the two views temporarily or spatially but the coding efficiency is not very good. MVC extension of H.264/AVC extends the current framework of H.264/AVC instead of using the computer vision (CV) paradigm to exploit the correlations between views. Block-based disparity compensated prediction (DCP) is adopted for inter-view prediction due to its similarity to motion compensated prediction (MCP). Many prediction techniques such as multiple reference frames (MRF) [4], variable block size (VBS) [5], sub-pixel MCP [6], hierarchical prediction structure [7], and fast motion estimation algorithms are already available for MCP. The differences between views are considered as camera panning from the one position to another one and the prediction error is encoded by transform coding. The major contribution of MVC extension is the Group Of Picture (GOP) structure that provides efficient DCP [8-9].

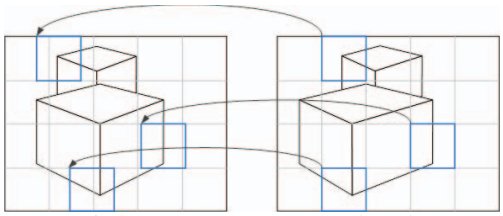

Fig. 1. Block matching of inter-view prediction in MVC

Although the RD performance is better than simulcast [10], the disparity model is still limited to block based translation and does not match the actual case very well. Fig. 1 shows a block matching of inter-view prediction. The block based differences between views are not simply translational.

To tackle the deformations between views, mesh based methods [11-12] were proposed for transforming a view to another. The prediction accuracy is improved by adopting the deformations formed by disparity effects but the complexity of handling the mesh is still high. Instead of generating a mesh, it is possible to approximate the deformations by providing prediction blocks or frames with various deformations. Among the deformation effects, Stretching, Compression and Shearing (SCSH) effects are the most common deformation between views, especially while the cameras are horizontally or vertically positioned. This approach was not very attractive in the past since it usually requires interpolation operation to obtain the deformed block or frames. Recently, a subsampled block matching technique [13] demonstrated an approximation of zoom motion compensated prediction in a low complexity way. By further generalizing the subsampled block matching idea, various types of deformations can be achieved by specially designed subsampling grid. In this work, SCSH DCP by subsampled block matching is proposed for inter-view prediction for MVC.

The paper is organized as follows. Section 2 gives a review of block matching techniques involved in MVC. The proposed SCSH techniques are presented in Section 3. Experimental results are given in Section 4 to reveal the improvements given by proposed techniques and the conclusion will be drawn in Section 5.

II. Stereo and Multiview Video Coding

2.1. Stereo and Multiview Video Coding

The major difference between MVC encoder and H.264/AVC encoder is the coding structure. Hierarchical coding used to form an efficient prediction structure for stereo video coding is shown in Fig. 2, in which I frames are only available in the left view. Black arrows indicate conventional inter frame prediction. Blue arrows indicate inter-view prediction. Dotted blue arrows are optional inter-view prediction. Inter-view prediction is used to remove the redundancies among different views.

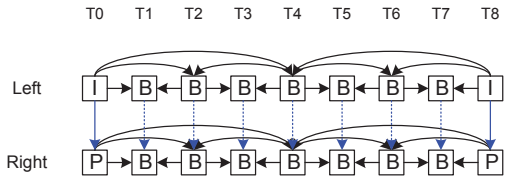

Fig. 2. Prediction structure of stereo video coding

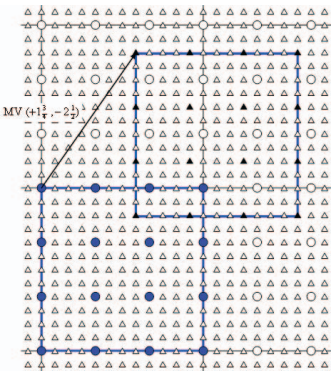

Fig. 3. Block-matching of conventional motion compensated prediction.

2.2. Block Matching Motion Compensated Prediction

In MCP, a frame is divided into non-overlapping blocks. Motion estimation is applied to find a prediction for each block based on the data in previously encoded frame. A residue block is created by subtracting the prediction from the current block. Only the residue block and the motion vector required to reproducing the prediction are encoded. The compression performance highly depends on the prediction accuracy. In H.264/AVC, several MCP tools are adopted to improve the prediction accuracy. Sub-pixel MCP enables more accurate motion vector up to 1/4 pixel precision. Fig. 3 illustrates the basic idea of sub-pixel MCP. The block for matching is obtained from the interpolated frame. Furthermore, with multiple reference frame technique, MCP can reference any frame over a long period of time so that the problem of temporary occlusion can be solved.

2.3. Block Matching Disparity Compensated Prediction

In stereo and multiview video coding, the frames capture the same scene at the same time with different camera locations. The correlation between views is very similar to single view video sequence with motion parallax effect. The difference between views depends on disparity effects. If the disparity information can be exploited like motions in MCP, the coding efficiency of the alternative views can be improved significantly. H.264/AVC MVC extension handles disparity compensated prediction (DCP) using the same set of coding tools for single view encoding. The reference frame from other views, instead of previous frames from the same view, is used in DCP. Practically, there is no additional parameter in the encoded bit-stream. The reference frame parameter indicates the inter-view frame and the motion vector parameter holds the disparity vector.

2.4. Limitation of Block-Matching based Disparity Compensated Prediction

The conventional DCP is based on block-matching assuming a translation motion model in which the disparity vectors of all pixels in a block are the same. However, the disparity model assumes each pixel has different disparity vector as the depth of every pixel in the frame can be different. In Fig. 1, the projected shapes of the objects in two views have small differences because of the depth within the objects. A real world example is also provided in Fig. 4 that shows

Left view **Right view**

Fig. 4. A real world stereo image example. The 3-D effect can be viewed by parallel eyes method.

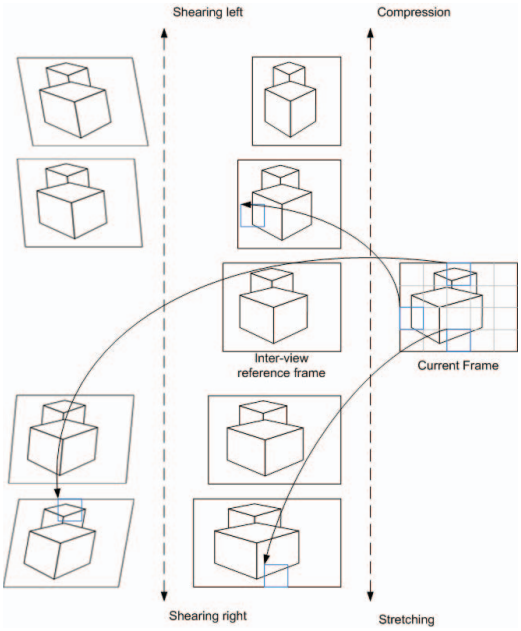

Fig. 5. Frame based SCSH disparity estimation.

the pixel based disparity model. In particular, vertical objects (e.g. walls) appear to be horizontally stretched or compressed between views. The horizontal objects (e.g. ceiling and floor) appear to be sheared between views. Based on this observation, it is possible to combine block-based approach with SCSH effects to provide the effects of pixel based disparity model. Although SCSH DCP can be achieved intuitively by a simple frame based approach as shown in Fig. 5, the complexity and the memory requirement of generating these SCSH frames make it impractical.

2.5. Subsampled Block Matching for MCP

Although SCSH effects can be achieved by applying affine transforms or by providing reference frames with SCSH effects, the computational complexity and the memory requirement are significant. Subsampled block-matching is recently proposed to efficiently provide zoomed reference frames for zoom motion compensated prediction. It subsamples the interpolated frame, which is already available for subpixel MCP, with various subsampling rates to obtain candidate prediction blocks with different zoom effects. It does not require additional operation to obtain a zoomed block nor additional memory space for storing zoomed frames and is successfully applied in Block-matching Translation and Zoom MCP (BTZMCP) [13]. The MCP can be generalized to include zoom reference frames such that it can better model the real world situation in which the projection of different regions or objects of a scene onto the imaging plane may exhibit zoom effects of various degrees. Fig. 6 shows an example of obtaining a 4/3-times zoomed block from the interpolated frame.

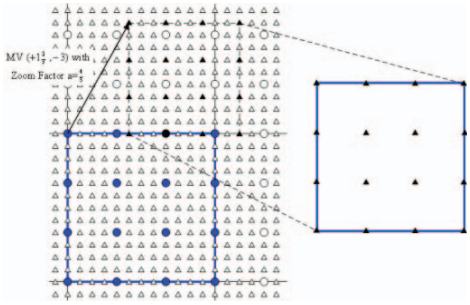

Fig. 6. Block-matching on a reference frame of zoom factor a=4/3.

III. SCSH BY SUBSAMPLED BLOCK MATCHING

Inspired by subsampled block matching, we can use different subsampling patterns to achieve more affine transformation effects. For quarter pixel MCP, the subsampling grid can be obtained by the transformation,

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 4 & 0 & u \\ 0 & 4 & v \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \quad (1)$$

where (x, y) and (x', y') are the relative coordinates of the pixels of in the current block and reference block, respectively. (u, v) is the translational motion vector in the interpolated frame. The grid is shown in Fig. 3. To provide zoomed candidate block, the subsampling factor is introduced into the transform matrix and the subsampling grid of BTZMCP becomes

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s & 0 & u \\ 0 & s & v \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (2)$$

where s = *(1, 2, ..., M)* is the subsampling rate associated with the zoom levels and the possible zoom scale are 4/s. With s=3, the zoomed block as shown in fig. 6 can be obtained.

SCSH DCP by subsampled block matching is proposed for inter-view prediction. Unlike BTZMCP, in which the subsampling rates are the same in both row and column directions, the subsampling grids of SCSH are asymmetric. Stretching and compression (SC) has only the horizontal sub-sampling rate changed. The subsampling grid of SC is defined as

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} sc & 0 & u \\ 0 & 4 & v \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (3)$$

where *sc = (1, 2, ..., M)*. The subsampling grid is illustrated in Fig. 7 and 8. Stretch and compression can be achieved without performing additional interpolation. Shearing (SH) can be obtained by the transformation

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 4 & sh & u \\ 0 & 4 & v \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (4)$$

where s*h = (-H, ..., -1, 0, 1, ..., H)* is the shearing factor that shifts the *x* coordinate depending on *y*. Shearing factor can be negative or positive such that the shearing can be left or right. Fig. 9 illustrates an example of subsampling grid of shearing.

It should be noted that the transform is applied on the subsampling grid instead of the reference frames. Thus, there is no transformation and interpolation operations involved if the resulting grid is hard coded in the codec. The overhead involved are: (i) the bits for indicating the SCSH parameter, which can be integrated with the reference frame number like BTZMCP, and (ii) a flag for indicating SCSH is on or off in the macroblock, which can be integrated with the block mode number. In addition, if the camera positions are up and down instead of left and right, the SCSH effects should be vertical instead of horizontal.

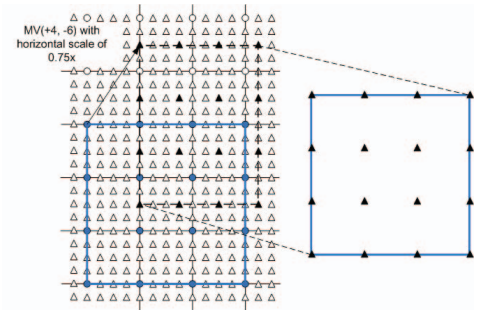

Fig. 7. Block-matching on a reference frame of compression factor of 4/3.

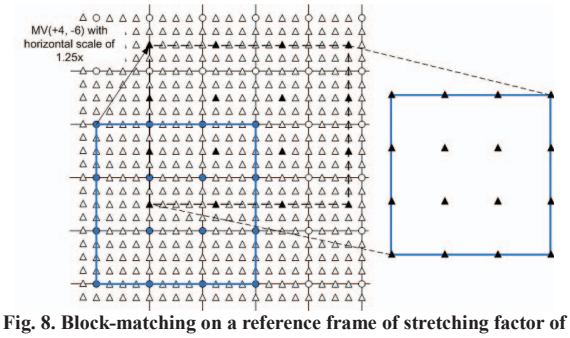

Fig. 8. Block-matching on a reference frame of stretching factor of 4/5.

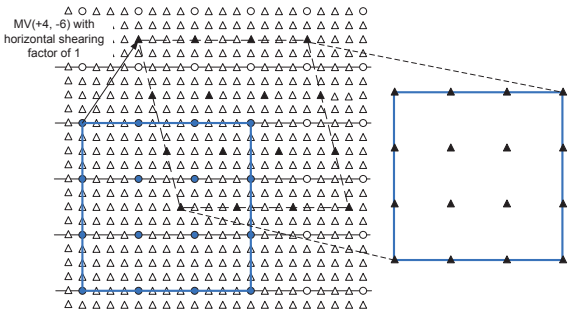

Fig. 9. Block-matching on a reference frame of horizontal shearing factor of 1.

IV. PERFORMANCE OF SCSH FOR INTER-VIEW PREDICTION

The inter-view prediction gain of SCSH by sub-sampled block matching will be presented via several experiments. Firstly, the direct improvement of SCSH will be compared to the conventional block based inter-view prediction approach. Secondly, the improvement of SCSH in commonly used MVC configuration is also provided to show the effect of SCSH in practical use.

4.1. Experiment setup

SCSH method is implemented on JM version 17.0 [14], which is the first version of JM with MVC support. SCSH is applied on large block modes (16x16, 16x8 and 8x16) of P frames only. In the experiments, eight stretching and compression and eight shearing candidates are used in DCP. Four sequences *ballroom, exit, vassar*, and *rena* used in JVT for developing H.264 MVC extension will be used and the sequences are in VGA (640x480) resolution. Each sequence has many views and two consecutive views are taken as a stereo pair. The first 100 frames from each view will be used. The H.264/AVC coding tools like VBS and RDO are turned on. Search

Table I - RD comparison of inter-view prediction between JM17 and JM17 with SCSH

vassar	JM17		SCSH	
QP	Bitrate	PSNR	Bitrate	PSNR
22	3200.57	41.63	3174.98	41.62
27	1439.78	38.11	1376.43	38.05
32	519.77	35.25	497.72	35.24
37	197.87	32.86	187.19	32.90
Average bitrate reduction (%)				-3.22
Average PSNR improvement (dB)				0.10

exit	JM17		SCSH	
QP	Bitrate	PSNR	Bitrate	PSNR
22	1851.47	42.32	1791.56	42.26
27	735.15	39.75	715.80	39.75
32	318.93	37.53	311.02	37.57
37	161.28	35.26	157.17	35.34
Average bitrate reduction (%)				-3.09
Average PSNR improvement (dB)				0.09

ballroom	JM17		SCSH	
QP	Bitrate	PSNR	Bitrate	PSNR
22	2930.93	41.90	2915.82	41.89
27	1463.84	38.83	1446.14	38.83
32	686.72	35.69	668.22	35.68
37	336.07	32.75	324.35	32.77
Average bitrate reduction (%)				-1.89
Average PSNR improvement (dB)				0.08

rena	JM17		SCSH	
QP	Bitrate	PSNR	Bitrate	PSNR
22	804.79	46.79	773.42	46.76
27	467.26	43.69	443.31	43.65
32	215.20	39.70	203.41	39.70
37	89.99	36.45	87.77	36.56
Average bitrate reduction (%)				-4.84
Average PSNR improvement (dB)				0.24

Table II - Comparison of overall RD performance between JM17 and JM17 with SCSH

vassar	JM17		SCSH	
QP	Bitrate	PSNR	Bitrate	PSNR
22	1612.61	38.733	1608.27	38.733
27	479.85	36.437	473.49	36.434
32	188.84	34.736	182.25	34.72
37	84.86	32.803	80.28	32.797
Average bitrate reduction (%)				-2.04
Average PSNR improvement (dB)				0.04

exit	JM17		SCSH	
QP	Bitrate	PSNR	Bitrate	PSNR
22	1011.55	40.181	1010.21	40.181
27	367.99	38.505	364.31	38.508
32	173.97	36.526	171.58	36.53
37	97.41	34.269	96.16	34.301
Average bitrate reduction (%)				-1.31
Average PSNR improvement (dB)				0.03

ballroom	JM17		SCSH	
QP	Bitrate	PSNR	Bitrate	PSNR
22	2062.89	39.466	2066.36	39.463
27	963.82	37.144	961.63	37.147
32	477.88	34.336	472.5	34.326
37	252.09	31.416	246.5	31.406
Average bitrate reduction (%)				-0.72
Average PSNR improvement (dB)				0.03

rena	JM17		SCSH	
QP	Bitrate	PSNR	Bitrate	PSNR
22	581.60	45.039	573.14	45.051
27	308.35	41.466	302.39	41.463
32	158.49	37.547	153.86	37.531
37	81.58	34.281	80.51	34.356
Average bitrate reduction (%)				-2.25
Average PSNR improvement (dB)				0.13

window is set at ±32 and exhaustive search is used within the search window. Left view is used as the base view and the right view is the alternate view predicted by inter-view prediction or inter prediction. Due to the special coding structure of MVC, P frames in right view use only inter-view prediction and B frames use only inter prediction. GOP structures without B frames and with 7 hierarchical B frames are tested. The average bitrate reduction and average PSNR improvement are calculated using Bjøntegaard's method.

4.2. Direct improvement of SCSH inter-view prediction

To investigate the direct improvement, GOP structure IIII is used for base view and PPPP for the alternate view. Since the P frames only use inter-view prediction, performance of SCSH and conventional block matching method can be compared directly. Table I shows the RD performance comparison of the alternate view from each sequence. The improvement is quite significant and the average bitrate reduction is around 1.89-4.84% and the average PSNR improvement is around 0.08-0.24dB.

4.3. Overall improvement of SCSH inter-view prediction

Practically, MVC uses prediction structures shown in fig. 2 that involved hierarchical B frames. However, inter-view prediction is normally not used as the inter prediction as bi-prediction already give very good predictions. As SCSH applies only on P frames, the improvement will be diluted by the B frames. In this part, the GOP structure as shown in fig. 2 is used, i.e. 7 hierarchical B frames between I and P frames. Table II shows that the RD performance of the alternate view that included all frames in that view. Although the improvement is diluted, it still have 0.72-2.25% of bitrate reduction and have 0.03-0.13dB of PSNR improvement.

V. Conclusion

A technique for inter-view prediction that exploit the transformation effect in stereo video is presented. Disparity compensated prediction in stereo and multiview video coding utilize frames from other views for inter-view prediction. With the conventional block-matching approach, the disparity effect within a block is limited to translation only. However, such assumption is not valid in reality. To better handle more realistic disparity effect like stretching, compression and shearing (SCSH) effects due to the depth of the objects, a new SCSH by subsampled block matching technique is proposed. It performs SCSH disparity estimation on the interpolated inter-view reference frame of the subpixel disparity estimation. Thus, no significant extra storage is required in both encoder and decoder implementation. In addition, the new SCSH disparity compensated prediction technique can be easily integrated into the multiview video codec to further improve the compression efficiency. The rate-distortion performance of SCSH is evaluated by implementation in JM reference software. Experimental results show that bitrate reduction up to 4.84% can be achieved in inter-view prediction. The improvement on the hierarchical prediction structure is up to 2.25%. Thus, subsampled block matching technique could be a promising direction in motion and disparity compensated prediction. In addition, the proposed subsampled block-matching on interpolated reference frames technique is also feasible for realizing other deformations with different camera array configuration.

Acknowledgment

The work described in this paper was substantially supported by a GRF grant with project number of 9041501 (CityU 119909) from City University of Hong Kong, Hong Kong SAR, China.

References

[1] "Introduction to Multiview Video Coding," ISO/IEC JTC 1/SC 29/WG 11 Doc. N9580, January 2008, Antalya, Turkey.

[2] ITU-T_and_ISO/IEC_JTC-1, "Generic coding of moving pictures and associated audio information – Part 2: Video," ITU-T Recommendation H.262 - ISO/IEC 13818-2 (MPEG-2), 1995.

[3] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 7, pp. 560–560, Jul. 2003.

[4] T. Wiegand, X. Zhang, and B. Girod, "Long-term memory motion compensated prediction," IEEE Trans. Circuits Syst. Video Technol., vol. 9, no. 2, pp. 70–84, Feb. 1999.

[5] G. J. Sullivan and R. L. Baker, "Rate-distortion optimized motion compensation for video compressionusing fixed or variable size blocks," in Proceedings of Global Telecommunications Conference, Phoenix, AZ, USA, 1991, pp. 85-90.

[6] T. Wedi and H. G. Musmann, "Motion- and Aliasing-Compensated Prediction for Hybrid Video Coding", IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 7, pp. 577–586, Jul. 2003.

[7] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF," in IEEE Int. Conf. Multimedia and Expo (ICME 2006), Toronto, ON, Canada, Jul. 2006.

[8] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient Prediction Structures for Multiview Video Coding," IEEE Trans. Circuits Syst. Video Technol., vol. 17, no. 11, pp. 1461-1473, Nov. 2007.

[9] M. Kitahara, H. Kimata, S. Shimizu, K. Kamikura, Y. Yashimata, K. Yamamoto, T. Yendo, T. Fujii, and M. Tanimoto, "Multi-view video coding using view interpolation and reference picture selection," presented at the IEEE Int. Conf. Multimedia and Exposition (ICME 2006), Toronto, ON, Canada, Jul. 2006.

[10] Y.J. Jeon, J. Lim, and B.M. Jeon, "Report of MVC performance under stereo condition," Doc. JVT-AE016, Joint Video Team, London, UK, June 2009.

[11] R.S. Wang and Y. Wang, "Multiview Video Sequence Analysis, Compression, and Virtual Viewpoint Synthesis, " IEEE Trans. Circuits Syst. Video Technol., vol. 10, no. 3, pp. 397–410, Apr. 2000.

[12] S. R. Han, T. Yamasaki, K. Aizawa, "Time-Varying Mesh Compression Using an Extended Block Matching Algorithm," IEEE Trans. Circuits Syst. Video Technol., vol. 17, no. 11, pp. 1506-1518, Nov. 2007.

[13] L.M. Po, K.M. Wong, K.W. Cheung, and K.H. Ng, "Subsampled Block-Matching for Zoom Motion Compensated Prediction", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 20, No. 11, pp. 1625-1637, Nov. 2010.

[14] "Joint Video Team (JVT) reference software version 17.0 [Online]. Available: http://iphome.hhi.de/suehring/tml/download/old_jm/."