EXPLOITING DEPTH INFORMATION FOR FAST MULTI-VIEW VIDEO CODING

Brian W. Micallef[1], Carl J. Debono[2] and Reuben A. Farrugia[3]

Department of Communications and Computer Engineering, University of Malta, Msida, Malta
{[1] brian.micallef, [2] c.debono}@ieee.org, [3] reuben.farrugia@um.edu.mt

ABSTRACT

Multi-view video coding exploits inter-view redundancies to compress the video streams and their associated depth information. These techniques utilize disparity estimation techniques to obtain disparity vectors (DVs) across different views. However, these methods contribute to the majority of the computational power needed for multi-view video encoding. This paper proposes a solution for fast disparity estimation based on multi-view geometry and depth information. A DV predictor is first calculated followed by an iterative or a fast search estimation process which finds the optimal DV in the search area dictated by the predictor. Simulation results demonstrate that this predictor is reliable enough to determine the area of the optimal DVs to allow a smaller search range. Furthermore, results show that the proposed approach achieves a speedup of 2.5 while still preserving the original rate-distortion performance.

Index Terms — 3DTV, disparity vector estimation, geometric disparity vector predictor, multi-view video coding.

1. INTRODUCTION

Multi-View Videos (MVVs) are formed by simultaneously capturing a scene from multiple cameras. Transmission of this data can be exploited to develop new applications, like 3D television (3D-TV) and Free-Viewpoint Videos (FVV). While 3D-TV offers the depth impression, FVV allows for interactive selection of viewpoint within a certain limited range. A new video format enabling both 3D-TV and FVV uses MVVs with their associated per-pixel depth data (N-video plus N-depth) [1], [2]. These videos are suitable since they allow view synthesis and rendering [3].

Multi-view Video Coding (MVC) is used for efficient compression of these videos, by exploiting spatial, temporal and inter-view redundancies [4]. This can be used to compress both color and depth data, where the latter can be considered as a luminance only video signal [5]. Although inter-view prediction improves the coding efficiency of a MVV, it also significantly increases the computational cost. This occurs because the inter-view redundancy is removed by conducting a technique similar to the Motion Estimation (ME) method across different views to obtain inter-view Disparity Vectors (DVs). This technique is generally the most time-consuming component in conventional video encoders [6]. Thus, an efficient Disparity Estimation (DE) technique is highly desirable for the encoder having a hybrid temporal/inter-view prediction structure [7]. An advantage of the MVVs is that DVs are highly dependent on the multiple camera setup and an estimate of the objects' depth. Both of these are easily obtained since the multi-view capturing system is normally precisely calibrated and the depth maps are usually available, since they are needed for view synthesis. However, if these are not available, they can be obtained from the same MVV [8] or captured by a depth map camera.

This paper proposes an efficient DE technique based on a reliable DV predictor. This is obtained using the multi-view geometry and depth data to reduce the search space. This decreases the complexity required in finding good candidate DVs and accelerates the inter-view prediction without any significant change in the objective or perceptive video quality or bit-rate. This predictor can be used with both iterative and fast DV search methods. The method is still compatible with the H.264/MVC standard since only the DV search process is modified. The performance of this technique is investigated for both the color and depth data compression of different N-video plus N-depth sequences where a speedup gain of 2.5 is achieved.

The rest of the paper is structured as follows: Section 2 describes the MVC standard. Section 3 introduces the basic multi-view geometry while section 4 describes the proposed DV predictor. Section 5 gives a description of the simulation environment while the following section presents the simulation results. Finally, Section 7 provides a conclusion for this work.

2. MULTI-VIEW VIDEO CODING

The MVV standard exploits the inter-view similarities to achieve efficient compression. This extends the ME techniques used to obtain motion vectors between temporal frames to adjacent inter-view images. This forms a combined temporal/inter-view prediction structure [4], referred to as MVC, giving significantly better results compared to using the H.264/AVC [9] solution on each view [10]. The optimal motion and disparity vectors can be derived using a full-search estimation process within a search area and those that minimize the rate-distortion

978-1-4244-7133-1/10/$26.00 ©2010 IEEE

matching cost [11] are selected. However, in practical MVC implementations, sub-optimal searching mechanisms, such as the Diamond searching strategy [12], are usually implemented since they reduce the number of full pixel search points while maintaining good rate-distortion efficiency. Conventionally, the starting and the central vector of the search area, is called the predictor and this is the median of the neighborhood vectors. To increase the estimation accuracy, the selected vector is refined to sub-pixel accuracy using a limited range [6]. Finally, this vector is transmitted as a difference vector from the predictor.

3. MULTI-VIEW GEOMETRY

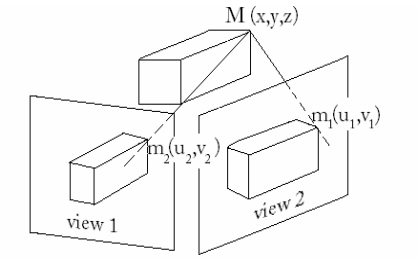

Fig. 1 Multi-view geometry

The advantage of a calibrated multi-view system is that an object can be easily located in all views by using multi-view geometry, as shown in Fig. 1. The matrix describing the linear mapping of a point $(u, v)^T$ to its corresponding 3-D point $(x, y, z)^T$ is called the camera projection matrix P and the equation can be easily determined by:

$$\zeta \boldsymbol{m} = P\boldsymbol{M} \tag{1}$$

where $\mathbf{M} = (x, y, z, 1)^T$ are the homogeneous coordinates of the 3-D point, $\mathbf{m} = (u, v, 1)^T$ are the homogeneous coordinates of the image point, and ζ is the distance of $\mathbf{M}$ from the focal plane of the camera referred to as the depth. The inverse of this equation gives the equivalent image point coordinates, of a 3-D point. The projection matrix P is a 3×4 full-rank matrix and is factorized as:

$$P = K[R \mid t] \tag{2}$$

where K is the camera calibration matrix, R is the rotation matrix and t is the translation matrix. The latter two parameters are known as the extrinsic parameters since they describe the orientation of the camera with respect to the external coordinate system. The camera calibration matrix K is dependent on the intrinsic parameters since it depends on the parameters within the same camera, such as the focal length f, image centre coordinates in pixels o_x, o_y and the pixel size in mm s_x, s_y along the two axes. This is given by:

$$K = \begin{bmatrix} f/s_x & 0 & o_x \\ 0 & f/s_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

Algebraically, each perspective view has an associated projection matrix which is generally provided for a calibrated camera system. All the image points which represent the same 3-D point from different views are called corresponding points [13].

4. PROPOSED DISPARITY ESTIMATION METHOD

An approximation of the DV predictor can be obtained by estimating the corresponding point in all the views. This can be done by exploiting knowledge of the geometrical location of an object within a target frame to search for the optimal disparity compensated MB of this object around the corresponding area in the reference frames. This region can be obtained by using the projection matrix of the target view, to virtually project the top left corner pixel location (m_1) of the sub-block to obtain a DV from the target view to a virtual 3-D point M. Then using the projection matrix of the respective reference frame, the corresponding points in these frames (m_2 and m_3) are located, as illustrated in Fig.2. The depth for this corner pixel is taken as the average depth of the whole sub-block, since this must represent the block's depth. This depth is obtained from the view's depth map, where a low resolution depth map was found to be sufficient. The predicted location in the reference frame is then subtracted from the current sub-block location, to obtain a translation vector to the new search area. This translation vector is an appropriate DV predictor that can be used to start the DV search in a smaller search area, as show in Fig. 2. A search area is still needed since the predicted DV does not necessarily return the DV which minimizes the distortion measure. However, since the optimal DV usually results in the vicinity of the predicted DV, we can reduce the search area size to a certain extent while still preserving the coding efficiency. This technique works only with the DE algorithms and it can be used to obtain a DV predictor for both the color and depth data. This occurs because the depth data of an object in a depth video can be located in the same way that color data of an object is located in a color video, thus this area can be used for compensation.

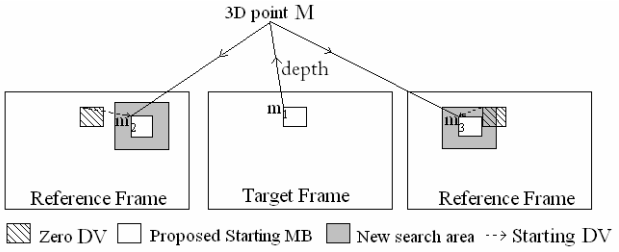

Fig. 2 The proposed search area

5. SIMULATION OVERVIEW

The proposed DE algorithm was implemented within the Joint Multi-view Video Coding Model (JMVC ver. 6.0) [14] to demonstrate its effectiveness. The JMVC model is provided to demonstrate the usefulness of the non-normative encoding techniques of the H.264/AVC standard, for MVC [15]; as defined in Annex H [9].

The JMVC encoder software was modified to locate the corresponding points of the top left pixel location, belonging to the current sub-block, in all the reference

frames. From this point, a translation vector from the current sub-block location is estimated. This is used as a DV predictor to start the optimal DV search in a smaller search area, using the original search estimation methods. If a corresponding point of this corner pixel falls outside the picture, then the median DV of the neighborhood is used. This technique was only implemented for DE while the original predictor and the whole search area are still used for ME. The proposed algorithm was implemented on all the important modes (16×16, 16×8, 8×16 and 8×8). For the other modes, the median DV obtained from the latter modes provides a very good DV predictor. In this technique, the depth map of the frame currently being encoded and the view's camera calibration parameters are only used to determine the corresponding points of the sub-block. These do not need to be transmitted since the optimal DV is still transmitted as a difference vector from the median DV, to produce H.264-MVC compatible bit-streams. The camera calibration parameters are fully compatible with the MVC supplemental enhancement information [16], so these can be transmitted.

Two MVVs with their associated depth data, known as the *Ballet* and *Breakdancers*, were used. These sequences are captured by eight cameras (1024×768, YUV 4:2:0, 15Hz) arranged on an arc with precise camera parameters [17]. For these simulations, the first three views, with a total of 100 frames each, were compressed. The Multi-view High profile was used to configure the encoder. The Group of Pictures (GOP) value of 1 was selected to get an encoding sequence of I-P-P-P with one temporal reference frame. The sequential inter-view prediction structure was defined such that all frames of view 2 are predicted from view 0 and all frames of view 1 are bi-predicted from both view 0 and view 2, to increase the coding efficiency [7]. The CAVLC was selected as the entropy encoder, to ensure further low delay characteristics. To allow random access [7], an Intra-coded frame was inserted every 12 frames. For the original DE and for ME, a search range of ± 32 pixels was used. A large search range had to be selected since the DVs are large in depth discontinuities. For the proposed DE algorithm, a smaller search range of ± 10 pixels was chosen to maintain almost the same rate-distortion performance. The estimation resolution is to quarter-pixel accuracy. Three different quantization parameters (QPs), 28, 30, and 32, were used to compare the rate-distortion performance. The parameters were chosen for a low complexity encoder, suitable for low delay applications. This technique uses the Full Search Estimation (FSE) or the Fast Search Estimation (FASE) [12] algorithms to determine the optimal disparity vectors and it is used to compress both color and depth videos.

All the simulations were carried out on a computer with an Intel® Core™ i7 processor with 12GB of RAM. The efficiency of the proposed DE algorithm was determined by the gain in speed obtained, when compared with the original estimation algorithms.

6. RESULTS AND ANALYSIS

Tables 1 and 2 represent a comparison of the MVC results obtained when encoding three views of the color and depth data of the *Ballet* sequence, respectively. These compare the performance obtained by the MVC when it uses the proposed disparity estimation technique with the FSE and FASE algorithms and when it uses the original DE with the FASE. These results represent the change in performance obtained from the original MVC with the FSE algorithm, since this gives the best prediction quality with the largest complexity. This comparison is in terms of the change in the average Peak Signal-to-Noise Ratio (PSNR), the percentage increase in total bit-rate, and the gain in speed.

Table 1. Comparison of the proposed and the original MVC performance on the color data of the *Ballet* sequence.

QP	FSE	Change	Prop. FSE	FASE	Prop. FASE
28	40.92 dB	Δ PSNR (dB)	-0.007	-0.008	-0.021
	1251.56 kbps	Δ Bit-rate (%)	+1.50	+0.66	+1.00
	66.53 hrs	Δ Speed	+2.40	+11.87	+21.56
30	40.17 dB	Δ PSNR (dB)	-0.014	-0.013	-0.022
	978.88 kbps	Δ Bit-rate (%)	+1.19	+0.64	+1.01
	66.51 hrs	Δ Speed	+2.41	+12.23	+21.97
32	39.33 dB	Δ PSNR (dB)	-0.041	-0.019	-0.033
	776.61 kbps	Δ Bit-rate (%)	+1.23	+0.35	-0.12
	66.38 hrs	Δ Speed	+2.40	+12.73	+22.07
Avg.	40.14 dB	Δ PSNR (dB)	-0.021	-0.013	-0.025
	1002.35 kbps	Δ Bit-rate (%)	+1.31	+0.55	+0.64
	66.47 hrs	Δ Speed	+2.41	+12.28	+21.87

Table 2. Comparison of the proposed and the original MVC performance on the depth data of the *Ballet* sequence.

QP	FSE	Change	Prop. FSE	FASE	Prop. FASE
28	46.33 dB	Δ PSNR (dB)	-0.069	-0.152	-0.164
	2036.13 kbps	Δ Bit-rate (%)	+0.87	+5.32	+5.59
	49.24 hrs	Δ Speed	+2.55	+10.84	+17.21
30	44.81 dB	Δ PSNR (dB)	-0.033	-0.172	-0.161
	1668.83 kbps	Δ Bit-rate (%)	+1.37	+4.76	+5.14
	49.17 hrs	Δ Speed	+2.55	+10.44	+18.24
32	43.29 dB	Δ PSNR (dB)	-0.085	-0.183	-0.188
	1375.90 kbps	Δ Bit-rate (%)	+1.56	+3.89	+4.79
	49.24 hrs	Δ Speed	+2.64	+10.64	+18.14
Avg.	44.81 dB	Δ PSNR (dB)	-0.063	-0.169	-0.171
	1693.62 kbps	Δ Bit-rate (%)	+1.27	+4.66	+5.18
	49.21 hrs	Δ Speed	+2.58	+10.64	+17.86

From these results, it can be determined that there is no significant change in the rate-distortion performance compared to the original algorithms and that the inter-view compression efficiency is retained. Thus, one can conclude that the proposed DV predictor is reliable enough to give a good estimate of the area where the optimal DVs are found. Moreover, the search area for DE can be drastically reduced, increasing the multi-view encoding speed. The proposed approach has achieved an overall speedup factor of 2.5 over the full-search process, while a speedup of 1.8 was achieved over the non-optimal diamond-search technique. These speedup gains are presented as an overall gain obtained when encoding the first three views of the MVV sequence. Encoding views with no inter-view

references, such as view 0, do not give any gain in speedup. Furthermore, the complexity of DE in MVC is proportional to the number of views with inter-view prediction and therefore the speedup gain is expected to increase when MVVs are encoded with more inter-view predicted views. Lower speed gains were registered with the FASE process because it has less full pixel search points. Smaller DV search areas can be used with the proposed DV predictor, giving a further increase in speed gain at the expense of some loss in rate-distortion performance.

Fig. 3 and 4 illustrate the PSNR versus bit-rate results for the color and depth data of the *Breakdancers* sequence, respectively. These results confirm that negligible loss in rate-distortion performance is obtained when using the proposed DV predictor with a reduced DV search area. The proposed method registered only a negligible average loss in PSNR of about 0.07dB. The gains in speed obtained for this sequence are similar to the ones in tables 1 and 2.

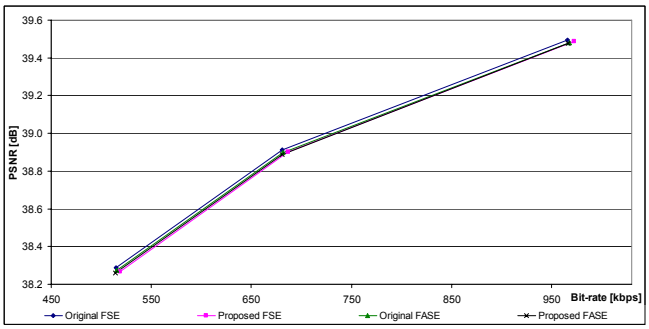

Fig. 3 Comparison results of the color data of the *Breakdancers* sequence.

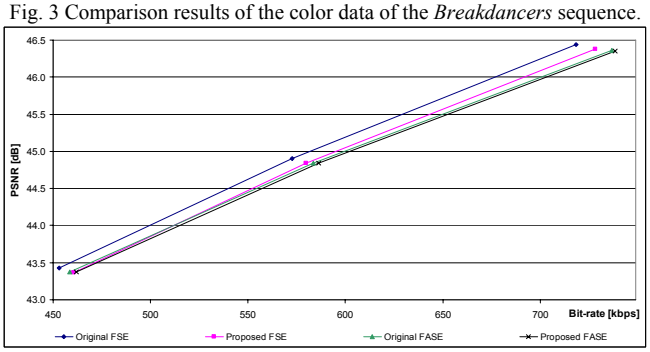

Fig. 4 Comparison results of the depth data of the *Breakdancers* sequence.

7. CONCLUSION

This paper presented a fast disparity estimation method for MVC based on a reliable disparity vector predictor, obtained using multi-view geometry. Experimental results have shown that this predictor is reliable enough to predict the position of the optimal disparity vectors that the search area can be reduced, significantly decreasing the encoding time. Speed gains of up to 2.5 were registered when encoding videos with three views and this gain increases further when encoding multi-view videos with more views. Simulation results have shown that these speed gains are achieved without affecting the rate-distortion optimization.

8. ACKNOWLEDGEMENT

This research work is partially funded by the Strategic Educational Pathways Scholarship Scheme (STEPS-Malta). This scholarship is partly financed by the European Union – European Social Fund (ESF 1.25). We would like to thank the Interactive Media Group of Microsoft Research for providing the *Breakdancers* and *Ballet* data sequences.

9. REFERENCES

[1] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O.Schreer, A. Smolic and R. Tanger, "Depth Map Creation and Image Based Rendering for Advanced 3DTV Services Providing Interoperability and Scalability," *Signal Processing: Image Communication. Special Issue on 3DTV*, February 2007.

[2] C. L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winderm and R. Szeliski, "High-Quality Video View Interpolation using a Layered Representation," *ACM SIGGRAPH and ACM trans. on Graphics*, pp. 600-608, Los Angeles, CA, USA, August 2004.

[3] C. Fehn, P. Kauff, M. Op de Beeck, F. Ernst, W. Ijsselsteijn, M. Pollefeys, L. Vangool, E. Ofek, and I. Sexton, "An Evolutionary and Optimised Approach on 3D-TV," *IBC 2002*, *Int. Broadcast Convention*, Amsterdam, Netherlands, September 2002.

[4] Y. Chen, Y. K. Wang, K. Ugur, M. M. Hannuksela, J. Lainema, and M. Gabbouj, "The Emerging MVC Standard for 3D Video Services," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, Article ID 786015, 13 pages, 2009.

[5] P. Merkle, A. Smolic, K. Müller and T. Wiegand, "Efficient Compression of Multi-view Depth Data Based on MVC," *ICIP 2007*, San Antonio, Texas, USA, pp 201-204, September 2007.

[6] M. E. Al-Mualla, C. N Canagrarajah, and D. R. Bull, *Video Coding for Mobile Communications, Efficiency, Complexity, and Resilience*, Elsevier Science, 2002, USA., pp. 93-200.

[7] ISO/IEC JTC1/SC29/WG11, "*Survey of Algorithms Used for Multi-view Video Coding (MVC)*," Doc. N6909, Hong Kong, China, January 2005.

[8] G. Zhang, J. Jai, T.-T. Wong and H. Bao, "Consistent Depth Maps Recovery from a Video Sequence," in *IEEE Trans. on Pattern Analysis and Machine Intell.*, vol. 31, no. 6, pp. 974-988, June 2009.

[9] ISO/IEC IS 14496-10, *Advanced Video Coding for Generic Audiovisual Services*, ITU-T Rec. H.264, March 2009.

[10] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient Prediction Structures for Multiview Video Coding," *IEEE Trans. on Circ. and Syst. for Video Tech.*, vol. 17, no. 11, pp. 1461-1473, November 2007.

[11] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-Constrained Coder Control and Comparison of Video Coding Standards," *IEEE Trans. on Circ. and Syst. for Video Tech.*, vol. 13, no. 7, pp. 688-703, July 2003.

[12] S. Zhu and K.-K. Ma, "A New Diamond Search Algorithm for Fast Block-matching Motion Estimation," *IEEE Trans. on Image Process.*, vol.9, no.2, pp. 387-392, February 2000.

[13] O. Schreer, P. Kauff and T. Sikora, *3D Video Communication: Algorithm, Concepts and Real-time Systems in Human Centred Communication*, John Wiley & Sons Ltd., England, 2005.

[14] ISO/IEC MPEG & ITU-T VCEG, "*Joint Multi-view Video Coding Model (JMVC 6.0)*," JVT-AE207, September 2009.

[15] ISO/IEC MPEG & ITU-T VCEG, "*MVC Software Manual*," JVT-AE207, September 2009.

[16] ISO/IEC MPEG & ITU-T VCEG, "*Revised Syntax for SEI Message on Multiview Acquisition Information*", JVT-Z038, January 2008.

[17] N-video plus N-depth data sequences with corresponding calibration parameters, from Interactive Visual Media Group at Microsoft Research, [Online]. Available: http://www.research.microsoft.com/vision/ImageBasedRealities/3DVideoDownload/

View publication stats