EXPLOITING DEPTH INFORMATION FOR FAST MOTION AND DISPARITY ESTIMATION IN 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 (MVC) employs both motion and disparity estimation within the encoding process. These provide a significant increase in coding efficiency at the expense of a substantial increase in computational requirements. This paper presents a fast motion and disparity estimation technique that utilizes the multi-view geometry together with the depth information and the corresponding encoded motion vectors from the reference view, to produce more reliable motion and disparity vector predictors for the current view. This allows for a smaller search area which reduces the computational cost of the multi-view encoding system. Experimental results confirm that the proposed techniques can provide a speed-up gain of up to 4.2 times, with a negligible loss in the rate-distortion performance for both the color and the depth MVC.

***Keywords*—**3DTV, fast disparity and motion estimation, geometric predictors, multi-view video coding.

1. INTRODUCTION

The color Multi-View Video (MVV) plus Depth (MVD) sequences are required to provide view synthesis [1] for efficient 3D [2, 3] and Free-Viewpoint Television [4] (3DTV/FTV) services. These applications require an enormous amount of bandwidth so various Multi-view Video Coding (MVC) schemes were developed to efficiently compress this information [5, 6]. The current state-of-the-art MVC scheme adopts also the Disparity Estimation (DE) to exploit the inter-view redundancies, apart from the spatial and temporal redundancies, to achieve a higher compression ratio. This coding scheme is efficient to compress both the color and the depth MVVs, where the latter can be considered as a luminance video signal only [6]. This coding efficiency is achieved with a drastic increase in the encoding time, since the estimation techniques are the most computational intensive components of the video encoding process [7]. Thus, efficient estimation techniques are desirable for MVC [8].

This paper proposes a technique that exploits the multi-view geometry and the depth information from the MVD sequence to obtain a higher fidelity Disparity Vector (DV) predictor. This information is used with the corresponding Motion Vectors (MVs) encoded for the reference view to obtain a better MV predictor for the Motion Estimation (ME) process. These geometric predictors are more accurate to determine the potential vectors requiring a smaller fixed search area in the estimation techniques, which significantly reduces the required computations. The performance of these techniques is investigated for both the color and the depth MVC of two different MVD sequences. Simulation results show that 76% of the encoding time can be saved for both video types.

The rest of the paper is structured as follows: Section 2 introduces the MVC standard and describes its complexity. Section 3 proposes the low computational motion and disparity estimation techniques. Section 4 gives the testing methodology used while section 5 presents the simulation results. Section 6 provides a conclusion for this work.

2. MULTI-VIEW VIDEO CODING

The current MVC scheme utilizes motion and disparity estimation techniques to remove the temporal and the inter-view redundancies, respectively. This coding scheme gives a significantly higher coding efficiency [5, 6] compared to simulcast coding. These estimation methods search for an optimal displacement vector to compensate each macroblock (MB) from either the temporal or view reference frame and then transmit only this vector for compensation. Conventionally, an optimal vector is exhaustively searched for each sub-MB partition, in all the search points within a fixed search area and in all the potential reference frames, for all the sub-MB combinations of the modes. For each search point, the Lagrangian Rate-Distortion (R-D) matching cost [5, 9] is computed and the optimal mode with its vectors that minimize the R-D cost are selected. The starting search vector, which is at the centre of the search area, is called the predictor and it represents the median of the neighborhood vectors. The selected vector is then transmitted as a residual vector from this predictor. This method is called the exhaustive Full Search Estimation (FSE) method. While this solution can give the optimal compensation vectors, its exhaustive search presents the most computationally expensive way to obtain them [7].

978-1-61284-162-5/11/$26.00 © 2011 IEEE

Fast Estimation (FASE) methods have been proposed in H.264/AVC to reduce the ME computation. Some of these sub-optimal techniques reduce the number of search points while still maintaining a good R-D efficiency [10]. These methods can also be used to speed-up the MVC. However, novel and more efficient techniques that exploit the multi-view geometry and the already encoded multi-view data to identify better the potential sub-MB replacements, to further speed-up the DE and the ME in MVC, need investigation.

3. PROPOSED DISPARITY AND MOTION ESTIMATION TECHNIQUES

The fundamental requirement of the DE is to obtain the optimal DVs that minimize the R-D cost. The DVs with the smallest distortion for a sub-MB are generally found in the vicinity of the sub-MBs corresponding areas in the reference views, which represent the same object. To identify these locations, the projection matrix P and the object's depth can be utilized. These are used to find the equivalent object's location in 3D space, and then re-locate it in the reference views, by using the multi-view geometry:

$$\zeta \boldsymbol{m} = P\boldsymbol{M} \quad (1)$$

where $\boldsymbol{m} = (u, v, 1)^T$ are the homogeneous coordinates of sub-MB's top-left corner, $\boldsymbol{M} = (x, y, z, 1)^T$ are the homogeneous coordinates of the 3D point, and ζ is the distance of $\boldsymbol{M}$ from the focal plane of the camera, referred to as the depth. This object's depth is estimated by averaging the pixel element (pel) values of the corresponding sub-MB in the depth video, of the current frame in the current view ($depth_t$). Since only the average of the sub-MBs is needed, low resolution depth maps, as those produced by the depth cameras, can be utilized. A translation vector from the zero DV to the identified corresponding point is formed and is used as a DV predictor to initiate the DV search area, as shown in Fig. 1. The search area can be reduced since this predictor gives a very good estimate of the optimal DVs. A similar method was already reported in our previous work [11] and can still be used to reduce the DE's computations.

The MVV sequence is representing the same moving objects within a scene from different views, thus, a strict relationship exists between their MVs. This relationship can be exploited to obtain an estimate of the MVs for the current view from the reference view, as illustrated in Fig. 1. This is done by first re-estimating the MV of the identified corresponding sub-MB, located by the new DV predictor, which has its MVs already encoded. The MV of this reference sub-MB is obtained by averaging the MVs enclosed within that sub-MB area. For the Intra coded blocks, the median MV is used since these contain no motion information. This MV displacement represents the optimal temporal sub-MB replacement to motion compensate the corresponding sub-MB in the reference view. By identifying the corresponding location of this temporal replacement in the current view, an estimate of the optimal sub-MB for motion compensation can be located. The depth map of the previous frame from the reference view ($depth_{t-1}$) is used to obtain these locations. A translation vector from the zero MV to the identified corner of this sub-MB is calculated. This gives an estimate of the optimal MV which can be used as a MV predictor to locate and reduce the MV search area. This technique can be used to estimate the motion within all the views in the MVV sequence, since appropriate MVs are estimated for each view. An analogy of this method is to obtain the equivalent 3D MV for the current sub-MB from the corresponding MVs in the reference view and the average depth change, and translate it to the current view to use it as MV predictor.

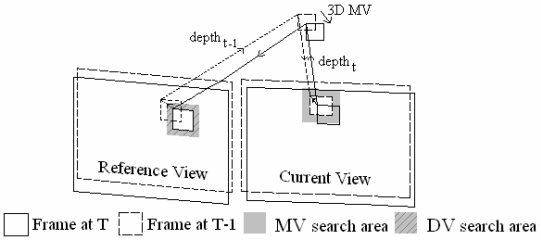

Figure 1: Motion and Disparity Estimation search areas

There are some exceptions when the geometric MV predictor should not be used. This occurs: 1) when the corresponding sub-MB in the reference view has a Zero MV, because it is more appropriate to consider the Zero MV as a predictor, since this represents a static object; 2) when the corresponding sub-MB falls outside the reference picture, the median MV must be used, since there is no way a MV can be estimated and 3) when encoding a view with no inter-view references; such as the Base view (view 0), since there are no reference MVs that can be used, thus, the median MV with the original search area must be utilized.

The optimal DVs and MVs derived using the proposed approach are still transmitted as residual vectors from their respective median vectors. The produced bit-streams thus remain in conformity with the H.264-MVC standard. This technique is designed to study the effect of the reduction in the computations used by the geometric estimation for fast encoding, while still aiming to maintain the original R-D performances and an unmodified decoder. However, if the depth MVV is transmitted for view synthesis, these predictors can also be used by the CODEC to reduce the rates of the optimal vector coding as demonstrated in [12]. Regarding complexity, the proposed estimation method differs from the original one by computing twice the corresponding points between different views using eqn. 1 and its inverse, and by computing the average MV of the corresponding sub-MB, which is insignificant compared to the whole MVC complexity. The camera calibration parameters are generally available for a calibrated system. Nonetheless, P can be estimated from the MVV in the absence of these parameters [13]. The camera calibration parameters are compatible with the MVC Supplemental

Enhancement Information [14], to ease their transmission. The proposed fast ME and DE techniques can be used for both the color and the depth MVC, since these contain the same multi-view properties.

4. SIMULATION OVERVIEW

The fast ME and DE strategies presented above were integrated within the Joint Multi-view Video Coding model (JMVC ver. 6.0) [15]. All the MB modes were enabled while the proposed algorithm was only implemented on the most frequently selected modes (16×16, 16×8, 8×16 and 8×8). The encoder was updated to record the time elapsed to encode the given MVV sequence. Finally, the original decoder was used to decode these bit-streams and MVV objective evaluation was performed.

Two MVDs, known as the *Breakdancers* and the *Ballet* sequence, were used [16]. The first three views of both the color and the depth MVVs with a total of 100 frames each were compressed. The Multi-view High profile was used to configure the encoder. Since the proposed low complexity estimation techniques are required for real-time applications, the simulation parameters were chosen to obtain a low complexity encoder. Thus, a Group of Pictures (GOP) value of 1 was selected to get an encoding sequence of I-P-P-P. 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 obtain the optimal inter-view coding efficiency [5, 8]. The CAVLC was selected as the main entropy encoder, to ensure low delay characteristics. To allow random access [8], Intra-coded frames were inserted every 12 frames.

A search range of ±32 pels was used for the original encoder while a smaller range of ±10 pels was chosen for the proposed technique. The estimation resolution is set to quarter-pel accuracy. Four different quantization parameters (QPs); 24, 28, 32, and 36, were used to compare the R-D performances [17-18]. The proposed geometric MV and DV predictors were used with the exhaust FSE and the diamond FASE [10] methods, to determine the optimal vectors for both the color and the depth MVC.

All the simulations were carried out on a computer with an Intel® Core™ i7 @ 3.20GHz CPU, with 6GB of RAM and running Microsoft Windows® 7 Ultimate x64. The efficiency of the proposed geometric estimations was determined as an encoding speed-up gain obtained by the modified encoder, when compared with the original one.

5. RESULTS AND ANALYSES

Tables 1 and 2 present the comparison results obtained when encoding the first three views of the color and the depth data of the *Ballet* sequence, respectively. These present the performance obtained by the MVC when it uses the proposed geometric estimation techniques with the FSE or the FASE method and when it uses the original estimation technique with the FASE method. These results are presented as the change in performance obtained from the original MVC using the exhaustive FSE method since this gives the best prediction quality with the largest complexity.

Table 1. The different R-D performances on the color data of *Ballet*.

QP	FSE	Change	Prop. FSE	FASE	Prop. FASE
24	42.04 dB	Δ PSNR (dB)	-0.006	-0.007	-0.011
	2348.43 kbps	Δ Bit-rate (%)	+0.40	+0.31	-0.25
	39.79 hrs	Gain in Speed	+4.12	+10.49	+26.15
28	40.92 dB	Δ PSNR (dB)	-0.018	-0.008	-0.020
	1251.56 kbps	Δ Bit-rate (%)	+0.77	+0.66	-0.14
	39.37 hrs	Gain in Speed	+4.06	+10.32	+26.61
32	39.33 dB	Δ PSNR (dB)	-0.051	-0.019	-0.052
	776.61 kbps	Δ Bit-rate (%)	+0.58	+0.35	-1.06
	38.82 hrs	Gain in Speed	+4.15	+10.67	+27.83
36	37.45 dB	Δ PSNR (dB)	-0.100	-0.032	-0.101
	521.98 kbps	Δ Bit-rate (%)	-0.19	+0.49	-2.13
	39.89 hrs	Gain in Speed	+4.07	+10.72	+28.38

Table 2. The different R-D performances on the depth data of *Ballet*.

QP	FSE	Change	Prop. FSE	FASE	Prop. FASE
24	48.97 dB	Δ PSNR (dB)	-0.059	-0.144	-0.179
	2948.38 kbps	Δ Bit-rate (%)	+1.20	+3.58	+3.70
	39.48 hrs	Gain in Speed	+4.48	+11.62	+25.11
28	46.33 dB	Δ PSNR (dB)	-0.108	-0.111	-0.171
	2036.13 kbps	Δ Bit-rate (%)	+1.39	+4.09	+3.59
	37.62 hrs	Gain in Speed	+4.26	+10.88	+24.57
32	43.29 dB	Δ PSNR (dB)	-0.127	-0.117	-0.182
	1375.90 kbps	Δ Bit-rate (%)	+1.09	+2.95	+2.74
	36.91 hrs	Gain in Speed	+4.19	+10.74	23.94
36	40.14 dB	Δ PSNR (dB)	-0.108	-0.106	-0.198
	882.17 kbps	Δ Bit-rate (%)	+1.48	+2.69	+2.79
	37.87 hrs	Gain in Speed	+4.34	+11.23	24.94

From these results, it can be deduced that the proposed estimation techniques give a significant speed-up factor of 4.2 times for the FSE and a speed-up factor of 2.5 times for the FASE, for both video types. The proposed technique registered only a negligible average loss of about 0.015dB and 0.12dB BD (Bjøntegaard Delta)-PSNR [17], from the quality provided by the original techniques, for the color and the depth data of this sequence, respectively. The original inter-view coding efficiency of about 15% BD-bit-rate saving for inter-view predicted views is also preserved.

When analyzing the results obtained by the MVC with both the proposed geometric estimations, with respect to using only the proposed DE in [11], it can be noted that the novel solution provides a further speed-up gain of 1.7 times for the FSE method and of 1.4 times for the FASE method. This speed-up gain was attained without any significant increase in the CODEC's complexity, thus, the latter is preferred. The main contribution of the encoder's speed-up is obtained when encoding the inter-view predicted views, since both the proposed ME and DE must be utilized by the MVC scheme. Thus, the speed-up gain is expected to increase when MVVs are encoded with more inter-view predicted views.

Fig. 2 illustrates the R-D results obtained for the color and the depth data of the *Breakdancers* sequence. Only a

negligible average loss of about 0.02dB and 0.11dB BD-PSNR, from the quality of the original algorithms, was registered for the color and the depth data of this sequence, respectively. The original inter-view coding efficiency of about 35% BD-bit-rate saving was also preserved. The gains in the encoding speed obtained for this sequence are similar to the ones presented in tables 1 and 2.

Fig. 3 shows that the original FSE algorithm with the smaller search area of ±10 pels fails to give the optimal compression and that it gives an increase in bit-rate of about 10%. It also illustrates that the basic multi-view encoding principle [5], with the static objects or with low dynamics being motion compensated and the objects with higher dynamics being disparity compensated, is also preserved for these sequences. In practice, the effective speed-up gain is doubled when both the color and the depth MVVs are encoded for the efficient 3DTV/FTV effect.

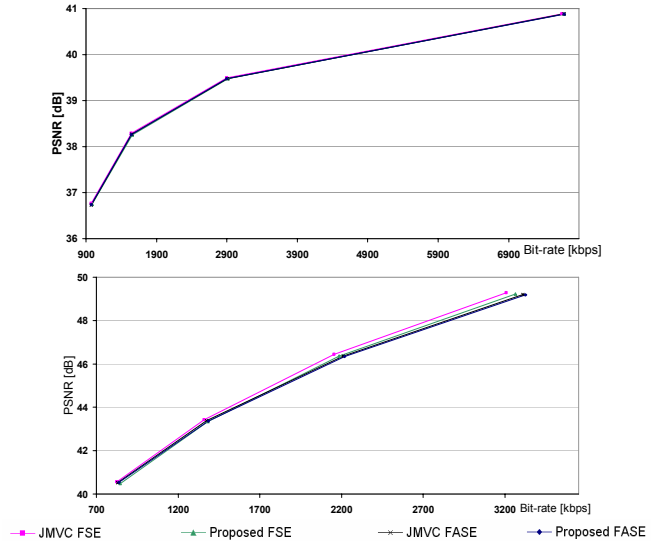

Figure 2: R-D curves for *Breakdancers* sequence (top: color, bottom: depth).


Figure 3: Frame 1 from view 2 of the *Breakdancers,* with black MBs being motion compensated and white MBs being disparity compensated, attained using the exhaustive FSE method with a QP of 32 (left-to-right: JMVM ±32 pels, JMVM ±10 pels, Proposed ±10 pels).

6. CONCLUSION

This paper presented a fast motion and disparity estimation technique that exploits the depth data of a MVD sequence together with the multi-view geometry, to obtain a more reliable motion and disparity vector predictor. Results have shown that these predictors are so accurate to determine the area of the optimal compensation vectors that the search area of the estimation techniques can be drastically reduced for both the full and the fast search estimation, attaining a significant gain in the MVV encoding speed. A speed-up gain of up to 4.2 times was registered when encoding the first three views of a MVV and this gain is expected to increase when MVVs with more inter-view predicted views are encoded. The results confirm that these gains were obtained for both the color and the depth MVC, without any significant change in the R-D performance or increase in the CODEC's complexity.

7. ACKNOWLEDGMENT

This research work is partially funded by STEPS-Malta and the EU-ESF 1.25. We would like to thank the Interactive Media Group for providing the MVD data sequences.

8. REFERENCES

[1] C. Zitnick, S. 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, Aug. 2004.

[2] 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 Comm. Special Issue on 3DTV*, Feb. 2007.

[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," in Proc. *IBC 2002*, Sept. 2002.

[4] ISO/IEC, "*Preliminary FTV model and requirements*," WG 11 Doc N8944, Jul. 2007.

[5] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient prediction structures for multi-view video coding," *IEEE Trans. on CSVT*, vol. 17, no. 11, pp. 1461-1473, Nov. 2007.

[6] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient compression of multi-view depth data based on MVC," in Proc. *ICIP 2007*, pp. 201-204, Sept. 2007.

[7] 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.

[8] ISO/IEC, "*Survey of algorithms used for multi-view video coding (MVC)*," Doc. N6909, Jan. 2005.

[9] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Trans. CSVT*, vol. 13, pp. 688-703, July 2003.

[10] 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, Feb. 2000.

[11] B. W. Micallef, C. J. Debono, and R. A. Farrugia, "Exploiting depth information for fast multi-view video coding," in Proc *PCS 2010*, Dec. 2010.

[12] B. W. Micallef, C. J. Debono, and R. A. Farrugia, "Exploiting depth information for efficient multi-view video coding," in Proc *ICME 2011*, Jul. 2011, *to appear*.

[13] Z. Zang, "Determining the epipolar geometry and its uncertainty: A re-view," *Int. Jour. on Comp. Vision,* vol 27, pp.161-195, Mar. 1998.

[14] ISO/IEC MPEG & ITU-T VCEG, "*Revised syntax for SEI message on multiview acquisition information,*" JVT-Z038, Jan. 2008.

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

[16] MSI sequences with calibration parameters [Online]. Available: http://research.microsoft.com/en-us/um/people/sbkang/3dvideodownload/

[17] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," Doc. VCEG-M33, Apr. 2001.

[18] Y. Su, A. Vetro, and A. Smolic, "Common test conditions for multiview video coding," Doc. JVT-U211, Oct. 2006.