Fast Inter-Mode Decision in Multi-view Video plus Depth 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*—Motion and disparity estimations are employed in Multi-view Video Coding (MVC) to remove redundancies present between temporal and different viewpoint frames, respectively, in both the color and the depth multi-view videos. These constitute the major computational expensive tasks of the video encoder, as iterative search for the optimal mode and its appropriate compensation vectors is employed to reduce the Rate-Distortion Optimization (RDO) cost function. This paper proposes a solution to limit the number of modes that are tested for RDO to encode the inter-view predicted views. The decision is based on the encoded information obtained from the corresponding Macro-block in the Base view, identified accurately by using the multi-view geometry together with the depth data. Results show that this geometric technique manages to reduce about 70% of the estimation's computational time and can also be used with fast geometric estimations to reduce up to 95% of the original encoding time. These gains are obtained with little degradation on the multi-view video quality for both color and depth MVC.**

Keywords—3DTV, geometric estimation, fast inter-frame mode selection, multi-view video plus depth, multi-view video coding.

I. Introduction

Multi-View Video (MVV) is a collection of videos that represents different viewpoints of the same scene. The data generated can be used to develop various services, such as Free-viewpoint Television (FTV) and Three-Dimensional Television (3DTV). For an enhanced visual experience, the display needs to create a viewpoint at any arbitrary position. Therefore, the corresponding depth MVVs must also be transmitted and used for Depth Image Based Rendering (DIBR) [1, 2]. This technique forms the Multi-view Video plus Depth (MVD) format [3], which drastically increases the bandwidth requirements, demanding more efficient MVV coding. Thus, the MVV Coding (MVC) standard was developed for efficient compression of both color [4] and depth [5] MVV types. This standard takes advantage of the fact that the inter-view redundancies can be exploited together with the temporal ones in a variable block sized hybrid Motion Estimation (ME) and Disparity Estimation (DE) technique. For more accurate and efficient compensation, the estimation process should be performed iteratively to compensate each partition of the available modes. These processes make the estimation the most computational intensive part of a video encoder [6], so more efficient techniques need to be developed for MVC to reduce this cost [7]. To the knowledge of the authors, exploitation of the geometrical information available from the depth data in MVD, for efficient and faster MVV coding of both the color and the depth MVVs, has only been considered in our previous work [8–11]. In these, we demonstrated that the multi-view geometry together with the available depth data can be used to identify better search areas for DE [9, 10] and ME [11]. Thus, these search areas can be reduced and so their associated computations, decreasing the computational demand with negligible influence on the color and the depth MVC efficiency. However, the estimation processes should still be performed repetitively on each mode's partition, to determine the optimal mode.

This paper proposes a solution that also uses the multi-view geometry together with the depth data to accurately identify the optimal corresponding Macro-block (MB) in the Base view. Then, it utilizes its encoded mode and motion vectors to determine the potential optimal modes for the currently being encoded MB, such that only the appropriate sub-optimal modes are tested for Rate Distortion Optimization (RDO). Results show that the technique can save up to about 70% of the encoding time required for the previous estimation techniques used in [9-11]. When compared with the original encoder, this solution together with the geometric estimations can save up to 95% of the original encoding time required to encode an inter-view predicted view, which finally results in an 84% reduction of the whole MVC time. These gains are obtained with a negligible effect on the MVV quality in both the color and the depth MVC of the two investigated MVD sequences.

The paper follows with section 2 describing the computational requirement of MVC and how this can be reduced. Section 3 proposes a solution for a fast inter-frame mode selection process that provides a further reduction in MVC computations. Then, section 4 gives the simulation overview used to determine the efficiency of the proposed method, while section 5 presents the simulation results obtained. Finally, section 6 concludes this work.

II. Multi-view Video Coding

The standard MVC method, originally presented in [4], is based on H.264/AVC. This states that the Decoded Picture Buffer of the variable block based ME can be extended to include also viewpoint frames as reference frames, to obtain a variable block based hybrid motion/disparity estimation. This imposes a substantial increase on the encoder's computational requirement since the Lagrangian RDO cost function [12] should be calculated exhaustively over all the search points in the temporal and viewpoint reference frames for ME [12] and DE [4], respectively, to obtain the optimal compensation vectors. Previous work demonstrated that the multi-view geometry together with the depth data can be used for MVD coding, to identify the corresponding inter-view replacements, allowing a reduction in the DE's search area [9, 10]. Further work showed that the motion information already encoded for this corresponding viewpoint location can be used to estimate 3D MVs and these can be used to locate and reduce also its ME search area [11]. These techniques can still be used to speed up ME and DE.

Nonetheless, for efficient R-D performance, the MB can be partitioned into one of the four main modes which are 16×16, 16×8, 8×16, and 8×8, where the 8×8 mode can be partitioned even further into four additional modes. This linearly increases the computational requirements since the ME and DE must be performed for each mode's partition, repetitively for all the modes, to obtain their optimal compensation vectors. Finally, the RDO cost function identifies the optimal mode together with its compensation vectors that should be transmitted for efficient MVC. Therefore, a method that identifies the potential optimal modes, to estimate only their compensation vectors and test only these modes, can further reduce the MVC computational requirement.

Some fast mode decisions for MVC are found in literature. The authors in [13] used the modes of the corresponding MB and its neighborhood in an encoded viewpoint, indicated by the global DV, to limit the modes tested for RDO. However, since the global DV is not always reliable, its neighborhood modes must be tested too. The authors in [14] utilize the epipolar geometry to find the candidate modes from another encoded viewpoint. However, more than one mode has to be tested since the optimal position along the epipolar line is unknown.

III. Proposed Inter-Frame Mode Selection

The multi-view geometry together with the depth data can be exploited to identify more accurately the corresponding encoded MB in an encoded viewpoint and then its encoded information can be utilized to limit the modes tested for the RDO of the current MB. The multi-view equation is given by:

$$\zeta \mathbf{m} = \mathbf{PM} \qquad (1)$$

where $\mathbf{M} = (x, y, z, 1)^T$ are the homogeneous coordinates of the 3D point, $\mathbf{m} = (u, v, 1)^T$ are the coordinates of the image point, $\mathbf{P}$ is the projection matrix that describes the linear mapping of the sub-MB's corner (u, v) from the Target frame to its corresponding 3D point (x, y, z), and ζ is the top-left sub-MB corner pixel depth value. Equation (1) is used to locate a 3D point in space for the currently being encoded MB and then its inverse is used to locate its corresponding encoded MB ($\mathrm{MB}_{\mathrm{corr}}$) in the viewpoint frame. This MB alone contains the optimal encoded information about the shape and dynamics of the same objects in the Base view. Thus, this helps to limit the modes tested for RDO in the currently being encoded MB to only the potential optimal ones. This method is based on the fact that the same objects in MVVs are captured from different viewpoints, thus high correlation exists among these areas and their modes in different views, and that the optimal ME modes are already encoded for the Base view [13][14].

Motion estimation is generally more efficient than disparity estimation to compress the MVVs, since generally the video is mostly static and there is more correlation between the temporal reference frames. So, during MVC the disparity/ motion estimation's RDO has a higher probability to use a partitioned mode only in the dynamic regions to describe better the dynamics and the boundary of the moving objects, as in H.264/AVC [15], and the optimal ME mode is encoded in the Base view. However, the disparity estimation is also efficient to encode the dynamic regions of the frame [4]. Thus, the partitioning of dynamic regions will also aid in finding a good compensation combination from the temporal and the viewpoint reference frames. However, for highly dynamic areas, the optimal mode combination for DE/ME is required since DE becomes more efficient. So, the mode selected by RDO in MVC should be dependent on the MB's dynamics.

Low complexity modes such as the INTRA, SKIP and 16×16 modes should always be tested first, since these do not require high computations and still provide a good R-D performance. Then, if the mode used to encode the corresponding MB is partitioned, this means that the current MB could be dynamic and this mode describes well the shape of the object. Therefore, for efficient encoding, the same mode should also be tested as a potential mode for RDO. To maintain coding efficiency, if the MB is encoded as an 8×8 mode, all the modes should be tested too. Then, if one of the coded optimal MVs for the corresponding MB is larger than ±3 pels, this area can be considered as being highly dynamic. If the corresponding MB is coded as an INTRA, this indicates that the ME in the Base view did not find a good temporal match due to three possible reasons: (1) area is so dynamic that it is more efficiently encoded as INTRA, (2) area contains very low texture, or (3) it is a newly exposed area, due to the objects' movement. Thus, if the corresponding MB of the current being encoded MB is encoded as INTRA or highly dynamic, all the modes should be allowed such that the optimal mode with the optimal combination of DE/ME from the viewpoint/temporal reference frames can be determined, for the current MB replacement. This is because the viewpoint reference frames can offer a better match for highly dynamic regions, since they can be more efficiently encoded using smaller DVs from the viewpoint reference frames [4]. Also, if the MB contains low texture or newly exposed areas, it can be disparity estimated since this data is already encoded for the Base view.

Given that only a small part of the image is dynamic or requires a complex mode, identifying these regions allows appropriate modes to be tested on them, leaving tests of only the simple 16×16 and SKIP modes on the rest of the image. These regions are identified from the already encoded optimal information for the corresponding MB in the viewpoint reference frame. Thus, generally, this method will result that the majority of the MBs in the frame are tested with only the SKIP and 16×16 modes, reducing a considerable amount of computations with minimum impact on coding efficiency.

For this technique, the optimally selected RDO modes for the Base view should be temporarily stored such that they are accessible while encoding inter-view predicted viewpoint videos. The mode limitation process discussed above is shown in Fig. 1.

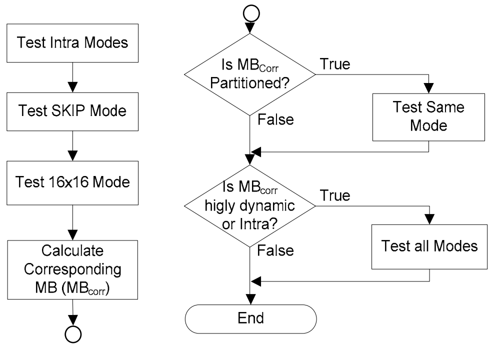

Figure 1. Flowchart of the Inter-Frame mode selection process.

IV. SIMULATION OVERVIEW

The proposed inter-frame mode selection algorithm was implemented within the Joint Multi-view Video Coding model (JMVC ver 6.0) [16] to evaluate its performance. The reference encoder was first modified to obtain the geometric DE as proposed in [9, 10] and the geometric ME as proposed in [11]. Then, it was further adapted to use the encoded information of the corresponding MB in the reference viewpoint, to limit the modes tested for RDO, as described above.

The *Breakdancers* and the *Ballet* calibrated MVD sequences were used to test the efficiency of the proposed algorithm [17, 18]. The first three views of both the color and depth MVVs of these sequences were encoded. Since fast estimation algorithms are more important in real-time applications, the simulation parameters presented in Table I were chosen to obtain a low complexity encoder. Both the Full Search Estimation (FSE) and the diamond FAst Search Estimation (FASE) [19] were used to determine the optimal compensation vectors for both the color and the depth MVC.

TABLE I. THE MVC SIMULATION PARAMETERS

Multi-view HIGH Profile
Real-time I-P-P-P temporal prediction structure [7]
Anchor frame period of 12
I – B – P for inter-view prediction structure
Entropy encoding with CAVLC
Original DE and ME with search area of ±32 pels [20] Geometric DE and ME with search area of ±10 pels [9, 11]
Estimation resolution of ¼ pel
Quantization Parameters (QPs) of 28, 32, 36, and 40

These simulations were carried out on a PC with an Intel® Core™ i7 @ 3.20GHz CPU, with 6GB of RAM and running Microsoft Windows® 7 Ultimate x64. The MVD videos were encoded with different MVC encoders and during each test the CPU encoding time was recorded. Then, the overall speed-up gains for the proposed encoders were estimated. Finally, the original decoder was used to decode the formed bit-streams and objective evaluation was performed.

V. RESULTS AND ANALYSIS

The R-D performance results obtained by the different MVC encoders are presented in Tables II and III. These present the overall performance obtained when encoding the first three views of the color and the depth data from the *Ballet* MVD sequence, respectively. The comparison consists of the original encoder, the encoder with the Geometric Estimations (GE) and the proposed encoder with both the geometric estimations and the Inter-Frame Mode Selection (IFMS) technique. Also, the performance obtained with either the FSE or the FASE, to find the optimal compensation vectors, are included within the results. The comparison is in terms of the percentage change in Peak-Signal-to-Noise Ratio (PSNR) in dB, the percentage increase in total MVV bit-rate in kbps, and the overall percentage reduction in encoding duration in hours, obtained with respect to the original encoder with FSE, which gives the optimal coding efficiency with the highest encoding time. Furthermore, Fig. 2 and Fig. 3 illustrate the R-D performance obtained by different MVC encoders on the color and depth MVVs of the *Breakdancers* MVD sequence, respectively.

TABLE II.
THE R-D PERFORMANCE FOR THE COLOR MVC OF THE *BALLET* SEQUENCE.

QP	Original with FSE	GE [9] with FSE	GE with IFMS and FSE	Original with FASE	GE [9] with FASE	GE with IFMS and FASE
28	40.92 dB 1251.56 kbps 39.37 hrs	-0.04% +0.77% -75.37%	-0.04% +2.53% -83.02%	-0.02% +0.66% -90.31%	-0.05% -0.14% -96.24%	-0.08% +2.01% -98.33%
32	39.28 dB 776.61 kbps 39.82 hrs	-0.12% +0.58% -75.90%	-0.18% +2.91% -82.73%	-0.05% +0.35% -90.63%	-0.13% -1.06% -96.40%	-0.18% +0.69% -98.39%
36	37.45 dB 521.98 kbps 39.89 hrs	-0.21% -0.19% -75.43%	-0.25% +1.57% -83.05%	-0.08% +0.49x -90.67%	-0.27% -2.13% -96.48%	-0.32% -0.52% -98.42%
40	35.30 dB 362.52 kbps 39.80 hrs	-0.28% +0.30% -75.12%	-0.30% +2.07% -83.79%	-0.13% -0.09x -90.88%	-0.30% -1.87% -96.73%	-0.38% -0.97% -98.55%

TABLE III.
THE R-D PERFORMANCE FOR THE DEPTH MVC OF THE *BALLET* SEQUENCE.

QP	Original with FSE	GE [9] with FSE	GE with IFMS and FSE	Original with FASE	GE [9] with FASE	GE with IFMS and FASE
28	46.33 dB 2036.13 kbps 37.62 hrs	-0.23% +1.39% -76.53%	-0.14% +2.68% -82.30%	-0.24% +3.09% -90.80%	-0.37% +2.59% -95.93%	-0.39% +3.26% -97.44%
32	43.29 dB 1668.83 kbps 37.59 hrs	-0.29% +1.09% -76.13%	-0.23% +2.98% -82.88%	-0.27% +2.95% -90.69%	-0.42% +2.74% -95.82%	-0.43% +3.19% -97.62%
36	40.14 dB 882.17 kbps 37.87 hrs	-0.27% +1.48% -76.96%	-0.21% +2.47% -82.52%	-0.26% +2.69% -91.09%	-0.48% +2.79% -95.99%	-0.48% +3.46% -97.99%
40	36.94 dB 528.96 kbps 39.05 hrs	-0.23% +1.65% -76.36%	-0.40% +3.18% -83.14%	-0.18% +1.22% -91.34%	-0.33% +0.61% -96.11%	-0.44% +2.60% -98.05%

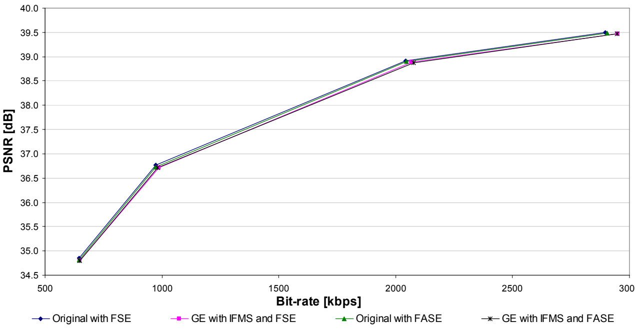

Figure 2. R-D curves for the color MVC of the *Breakdancers* sequence.

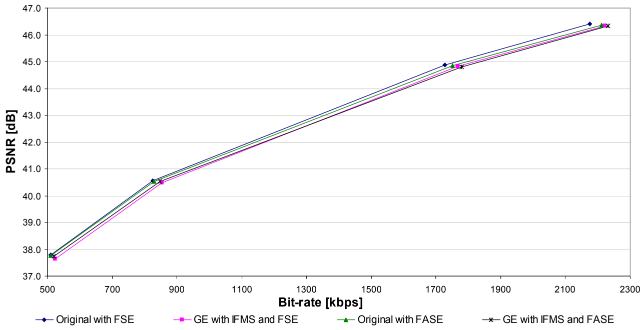

Figure 3. R-D curves for the depth MVC of the *Breakdancers* sequence.

The results obtained indicate that there is only a small average loss of 0.07dB Bjøntegaard Delta (BD)-PSNR [21] in the average color MVV quality and 0.09dB BD-PSNR in the average depth MVV quality, when the proposed geometric method is applied to limit the candidate modes for RDO. These results are obtained after averaging the R-D results obtained for each view and for both the tested sequences. Although a small loss in quality is registered, while encoding the inter-view predicted videos, the proposed technique gives an average speed-up gain of about 3.2 for FSE, which correspond to a reduction of 70% in its computational time. Furthermore, it gives a speed-up gain of 2.4 for FASE, which correspond to a reduction of 60% in its encoding time. This results in overall speed-up gains of up to 1.5 times for FSE and up to 2.3 times for FASE when considering the whole MVC with the Base view encoding, presented in the tables above. However, the speed-up gains achieved start to decrease slightly as the required video quality increases, as in the Base view more MBs are encoded with partitioned or INTRA coded modes. This in turn increases the number of modes tested for RDO of the inter-view predicted views which allows for the required increase in video quality for the inter-view predicted videos. When using both GE and IFMS methods, these on average reduce about 95% of the computational time for FSE and 85% for FASE while encoding an inter-view predicted view, which finally results in a speed-up gain of up to about 6.1 times for the whole MVC, for both FSE and FASE. This is achieved with only a small average quality loss of 0.11dB BD-PSNR in the color MVV quality and 0.18dB BD-PSNR in the depth MVV quality, when compared to the original encoder. Therefore, the proposed technique that limits the inter-frame modes for RDO demonstrated to be efficient and can also be used with the GEs to obtain an extremely fast encoder.

VI. Conclusion

A fast inter-frame mode selection process was presented in this paper. It utilizes the multi-view geometry together with the depth data to obtain the corresponding inter-view position. Then, the encoded mode and motion information for this location are used to limit the modes tested for Rate-Distortion Optimization of the MB currently being encoded in the inter-view predicted frame. Simulation results demonstrated that this technique can be used together with geometric estimations in order to achieve significant speed-up gains of up to 6.1 times, while imparting only negligible quality degradation in both color and depth MVC, when compared to the original encoder.

Acknowledgments

This research work is partially funded by STEPS-Malta and partially by the European Union – ESF 1.25. We would like to thank the Interactive Media Group of Microsoft® Research for providing the *Breakdancers* and *Ballet* MVD sequences.

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 Comm. Special Issue on 3DTV*, Feb. 2007.

[2] Y. Mori, N. Fukushima, T. Fujii, and M. Tanimoto, "View generation with 3D warping using depth information for FTV," in Proc of *3DTV Conference*, pp. 229-232, May 2008.

[3] ISO/IEC MPEG & ITU-T VCEG, "*Multi-view Video plus Depth (MVD) format for advanced 3D video systems*," Doc. JVT-W100, Apr. 2007.

[4] P. Merkle, K. Mueller, A. Smolic, and T. Wiegand, "Efficient compression of multi-view video exploiting inter-view dependencies based on H.264/MPEG-AVC," in Proc. of *ICME 2006*, Jul. 2006.

[5] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient compression of multi-view depth data based on MVC," in Proc. of *3DTV-Conference*, May 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 MPEG & ITU-T VCEG, "*Survey of Algorithms Used for Multi-view Video Coding (MVC)*," Doc. N6909, Jan. 2005.

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

[9] B. W. Micallef, C. J. Debono, and R. A. Farrugia, "Exploiting depth information for fast multi-view video coding," in Proc. of *PCS 2010*, pp. 38-41, Dec. 2010.

[10] B. W. Micallef, C. J. Debono, and R. A. Farrugia, "Fast disparity estimation for multi-view video plus depth coding," in Proc. of *VCIP 2011*, Nov. 2011.

[11] B. W. Micallef, C. J. Debono, and R. A. Farrugia, "Exploiting depth information for fast motion and disparity estimation in multi-view video coding," in Proc. of *3DTV-Conference 2011*, May 2011.

[12] 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 CSVT*, vol. 13, pp. 688-703, Jul. 2003.

[13] L. Shen, Z. Liu, T. Yan, Z. Zang, and P. An, "View-adaptive motion estimation and disparity estimation for low complexity multiview video coding," *IEEE Trans. on CSVT*, vol. 20, no. 6, pp. 925-930, Jun. 2010.

[14] G. Yang, L. Liang, and W. Gao, "An epipolar restricted inter-mode selection for stereoscopic video encoding," in Proc. of *PCS 2010,* pp. 338-341, Dec. 2010.

[15] H. Q. Zeng, C. H. Cai, and K. –K. Ma, "Fast mode decision for H.264/AVC based on macroblock motion activity," *IEEE Trans. on CSVT*, vol. 19, no. 4, pp. 491-499, Apr. 2009.

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

[17] MSR MVD Sequences [Online]. Available: http://research.microsoft.com/en-us/um/people/sbkang/3dvideodownload/

[18] 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.

[19] 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. 287-290, Feb. 2000.

[20] ISO/IEC MPEG & ITU-T VCEG, "*Common Test Conditions for Multiview Video Coding*," Doc. JVT-U211, Oct. 2006.

[21] G. Bjøntegaard, "*Calculation of Average PSNR Differences Between RD-Curves*," Doc. VCEG-M33, Apr. 2001.