Reducing 3D Video Coding Complexity through More Efficient Disparity Estimation

Brian W. Micallef, *Member*, IEEE, Carl J. Debono, *Senior Member*, IEEE, and Reuben A. Farrugia, *Member*, IEEE

Abstract — *3D video coding for transmission exploits the Disparity Estimation (DE) to remove the inter-view redundancies present within both the texture and the depth map multi-view videos. Good estimation accuracy can be achieved by partitioning the macro-block into smaller sub-blocks partitions. However, the DE process must be performed on each individual sub-block to determine the optimal mode and their disparity vectors, in terms of rate-distortion efficiency. This vector estimation process is heavy on computational resources, thus, the coding computational cost becomes proportional to the number of search points and the inter-view modes tested during the rate-distortion optimization. In this paper, a solution that exploits the available depth map data, together with the multi-view geometry, is proposed to identify a better DE search area; such that it allows a reduction in its search points. It also exploits the number of different depth levels present within the current macro-block to determine which modes can be used for DE to further reduce its computations. Simulation results demonstrate that this can save up to 95% of the encoding time, with little influence on the coding efficiency of the texture and the depth map multi-view video coding. This makes 3D video coding more practical for any consumer devices, which tend to have limited computational power*[1].

Index Terms — 3D video coding, fast disparity estimation, fast mode selection, disparity estimation, multi-view video coding.

I. Introduction

With the latest developments in advanced multimedia video systems, consumers' devices can support more immersive video services, such as the Three-Dimensional TV (3DTV) [1] and/or the Free-Viewpoint TV (FTV) [2]. These allow the users to experience a more realistic 3D scene representation with the potential of interactively selecting an arbitrary viewpoint from a restricted range. To effectively enable these technologies, 3D videos in the Multi-view Video plus Depth (MVD) format [3], which consists of the texture and the depth map Multi-View Video (MVV) data, need to be transmitted. The texture MVVs represent the scene from fixed viewpoints, while their corresponding depth map MVVs allow for Depth Image-Based Rendering (DIBR) of a virtual texture viewpoint, at any arbitrary position [4]. The depth map data can be generated in real-time, using dynamic programming on a General purpose Graphical Processing Unit (GpGPU) [5]. This creates a vast amount of MVV data that must be transmitted over limited bandwidth channels. As a result, efficient Multi-view Video Coding (MVC) techniques become crucial for the success of the 3D video (3DV) technology. The standardized MVC encoder is the H.264/MVC and is already used for stereo 3DV coding on Blu-ray™ media. Consequently, the next logical step is to transmit 3DVs over the latest transmission channels, such as the Digital Video Broadcasting (DVB) and Long Term Evaluation (LTE) networks, for consumers' consumption. This standard is also adequate since it provides backward compatibility with current HD-video coding and transmission systems using H.264/AVC [6]; allowing consumers using new devices to view 3D media in 3D, as intended [7], and those using old devices to still view the same content in 2D.

The H.264/MVC exploits the inter-view redundancies by extending the Motion Estimation (ME) technique of the H.264/AVC, to operate also across viewpoints' frames, to perform Disparity Estimation (DE). This results in a system that is more efficient than simulcast coding [6]. The DE inherits the search in multiple reference frames and varies block sizes from ME, which makes the ME process the most computational intensive component [8]. This brings higher coding efficiency at the expense of doubling the already high computational complexity to the system. These coding schemes demonstrated to be efficient to compress both the texture [9] and the depth maps of the 3DVs [10], and again this effectively doubles the 3DV coding's computational cost requirement, which hinders its implementation on common consumer devices. Therefore, to deliver such services within low-latency and at an affordable price, and to promote more applications for 3DVs, especially on consumer devices with limited power and computational resources [11], less computationally intensive algorithms are required [12].

Several fast algorithms that are specifically designed for DE in MVC, were proposed in literature. San *et al* [13] employed an epipolar-based fast DE algorithm that greatly reduces the search range by searching only around the epipolar lines where the optimal Disparity Vectors (DVs) should lie. Kim *et*

[1] This research work was partially funded by the Strategic Educational Pathways Scholarship Scheme (STEPS-Malta) and by European Union - European Social Fund (ESF 1.25).

B. W. Micallef is with the Department of Communications and Computer Engineering (CCE), University of Malta, Msida, MSD 2080, Malta (e-mail: brian.micallef@ieee.org).

C. J. Debono is with the Department of CCE, University of Malta, Msida, MSD 2080, Malta (e-mail: c.debono@ieee.org).

R. A. Farrugia is with the Department of CCE, University of Malta, Msida, MSD 2080, Malta (e-mail: reuben.farrugia@um.edu.mt).

Contributed Paper
Manuscript received 12/09/13
Current version published 03/20/14
Electronic version published 03/20/14.

al [14] utilized the geometry of the camera arrangements to determine the reliability of the DV and the Motion Vector (MV) predictors based on their relationship, and depending on the accuracy of both predictors, adaptively adjusted the search areas for both the DE and the ME, respectively. This correlation together with that among the cameras, is also used by Li *et al* [15] to further reduce these computations. Pan *et al* [11] simplified the method of obtaining the global DV and depending on the statistical correlation between the optimal DV, the global DV, and the DV predictor, they adaptively reduce the DE's search area. Zhu *et al* [16] first utilized the DVs encoded for the previous temporal frame to obtain a DV predictor. Then the search area center was selected from the spatial or the temporal neighborhood DVs and the search area was adapted depending on the distance between the two.

All these methods adaptively reduced the optimal DV search area according to the reliability of the DV predictor. However, the DE is still highly computational intensive as the optimal DV must be calculated for every partition of the seven different modes. Traditional ways were used to reduce the mode selection process for MVC [17]. Other MVC mode decision algorithms [18]-[23] adapted the optimal mode determined for the corresponding Macro-Block (MB) in the neighborhood encoded viewpoints. However, these are primarily based on the optimal mode determined from the Base viewpoint; which defines the optimal compensation between temporal frames and not the optimal inter-view one or their optimal combination. Nevertheless, mode decision in ME depends on the movement of the objects in the MB, while for the latter, it depends on the objects' depth in the MB and their disparity. Therefore, this optimal mode may be inefficient for DE, since the characteristics between viewpoints are not considered [11]. Consequently, a method to determine the specific optimal mode for DE is necessary. This mode can be used for anchor frame coding or tested with the optimal ME's mode during Rate-Distortion Optimization (RDO).

Since the depth map MVV data is rich in geometrical information about the scene, and is already available in 3DVs, it is wise to exploit this data to assist the 3DV encoding stage. Thus, this paper proposes a technique that first utilizes the depth map data together with the multi-view or the Homogenous equations to geometrically calculate a more accurate DV predictor, which allows a permanent reduction in the DE's search area. Then, the average depth map values of the basic partitions forming the main modes are used to determine the potential sub-optimal modes to test specifically for DE. To the authors' knowledge, no work in literature except those of the same authors [23]-[27] exploits the full geometrical information provided by the depth map data to reduce the computational burden of 3DV coding through such geometrical properties. Furthermore, this did not consider exploiting the depth map data to jointly reduce both the search area and the DE's modes, during RDO. Results demonstrate that an average speed-up gain of about 26 times was registered over the original DE, out of which a gain of 7.4 was obtained by the reduction in the search area, while a gain of 3.6 was achieved by the sub-optimal mode decision process. These gains were obtained for DE within both the texture and depth map MVC with negligible influence on their performance.

The rest of the paper is structured as follows: Section II discuses the H.264/MVC standard and the main computational complexity of its RDO process. Section III introduces the proposed MVC technique that exploits the depth map data to determine the position for the disparity's search area and the sub-optimal modes that can be tested for RDO, during DE. Section IV gives the testing methodology used, while section V presents the experimental results obtained. Finally, section VI provides a conclusion for this work.

II. Multi-view Video Coding

The spatial and the temporal redundancies present within the MVVs are removed as in conventional single-view video coding with the H.264/AVC, through the Intra coding and the ME techniques, respectively. Additionally, the H.264/MVC allows also the use of viewpoint frames, as reference frames in the encoder's Coded and the decoder's Decoded Picture Buffers [6], which are at the same temporal instance of the target frame, but from another encoded viewpoint. This is done to extend the built-in block-based ME to additionally perform also the block-based DE, to suppress in a similar way the inter-viewpoint redundancies. This MVC scheme is illustrated in Fig. 1 and is defined as an extension of the H.264/AVC standard, in its Annex H [28]. In doing so, this generates a significant increase in computational requirements.

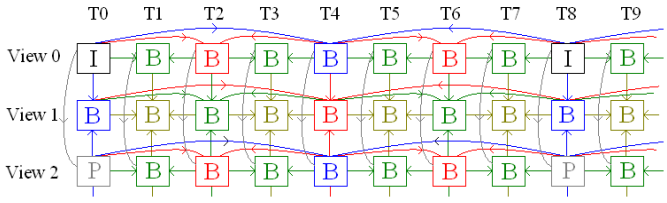

Fig. 1. The Multi-view Video Coding structure.

The ME and the DE techniques defined within the MVC standard utilize the block matching process to find an estimate of the current MB; where compensation vectors establish a correspondence between similar sub-MBs' partitions between the Target and the Reference frames, to indicate from where the current MB's parts can be compensated, instead of being re-encoded. For an optimal coding efficiency, an exhaustive search for the most efficient compensation vector, for every sub-MB partition within a mode, is performed in all the search points within a search area and iteratively for all the partitions within the available modes. This search is also performed in all the temporal and the viewpoint reference frames to find the appropriate partition's MV or DV, respectively. For an efficient compensation, the MB can be partitioned into one of the seven main modes illustrated in Fig. 2. The Lagrangian's RDO matching cost [29] function is computed every time, and the optimal mode with its optimal compensation vectors are defined such that they minimize this cost function. These are then transmitted together with the residual data as part of the bit-stream. The predictor defines the initial position

where the estimation technique should start searching for these optimal vectors, and it locates the center of this search area. According to the standard, this corresponds to the median of the neighborhood encoded vectors. Only the difference between the optimally selected vector and this prediction is transmitted to further improve the coding efficiency [30]. This method is called the exhaustive Full Search Estimation (FSE). To effectively reduce these computations, a sub-optimal Fast Search Estimation (FASE), such as the Diamond Search [31], can be performed. This systematically reduces the number of search points within the search area while still aiming to maintain almost optimal coding efficiencies.

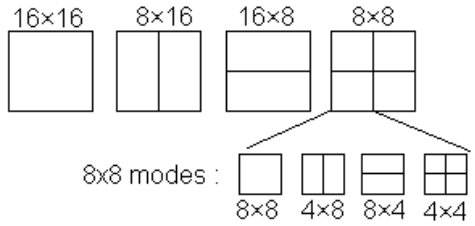

Fig. 2. The available inter-view modes used for Disparity Estimation [30].

For ME, the median MV predictor can indicate adequate potential compensation replacements because statistically the neighborhood vectors are highly correlated; as the majority of the MBs in a picture are usually static or with homogeneous movement. However, it can completely fail for DE, since the DVs tend to easily lose their neighborhood correlation with depth discontinuities or a large camera separation [9]. Due to this, a large search area has to be ensured, imposing a long DE's encoding duration [8]. Several faster DE methods adaptively adjust the size of their search area depending on the reliability of this DV predictor or the neighborhood disparity. However, this only indicates that a more reliable DV predictor, rather than the median one used in the standard, is required. If such a DV is used, better prediction of the potential optimal DVs is achieved and the search area around it can be reduced to a small fixed one, together with its computations, with minimal impact on coding efficiency. The computational cost of DE comes also from the fact that an iterative search process has to be performed while estimating each mode, to identify the optimal one for compensation. This refinement linearly increases the DE's computational cost and makes it proportional to the number of candidate modes tested for RDO. Thus, more accurate mode prediction will restrict this testing to only the potential optimal ones, further reducing the computational cost of the DE. However, for better prediction, this mode prediction process has to be designed specifically for the DE.

III. Proposed Disparity Estimation

During DE, the optimal mode and its sub-MBs' DVs should be highly dependent on the objects' depth and the cameras' setup, which are now readily available for transmission in the 3DVs. This is because the optimal DVs should lie around the corresponding areas, which do represent the same objects from different viewpoints [32], and thus, these similar blocks reduce the estimations' distortion. The multi-view geometry together with the depth map data can be utilized to perform this 3D point correspondence of the partitions' top-left corner, to identity the potential low-distortion replacement within the viewpoint reference frames, through:

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

where $\boldsymbol{m}=(u, v, 1)^T$ represents the coordinates of the image points, $\boldsymbol{M}=(x, y, z, 1)^T$ is the equivalent 3D point in space, $\mathbf{P}$ is the 3×4 projection matrix, and ζ is the referred to as the depth ($depth_T$) [32]. Equation (1) can be used to identify the current sub-MB's location $\boldsymbol{m_T}$ within the Target frame, to its 3D point $\boldsymbol{M}$, using the projection matrix $\mathbf{P_T}$. This can then be used again to relocate this equivalent partition's positions in the viewpoint reference frames $\boldsymbol{m_{R0}}$ and $\boldsymbol{m_{R2}}$, using their appropriate projection matrices $\mathbf{P_{R0}}$ and $\mathbf{P_{R2}}$ and their depths, as indicated in Fig. 3. These points can also be identified through the Homogenous equations; where a direct 2D correspondence can be determined for the current integral depth map level, using:

$$\boldsymbol{m_R} = \mathbf{H}_{depth_T}\boldsymbol{m_T} \quad (2)$$

where $\mathbf{H}_{depth_T}$ is the 3×2 Homogeneous matrix that depends on the current depth level; $depth_T$ [32]. An estimate of the objects' depth, required in both equations, is obtained by averaging the depth map pixel element (pel) values of the sub-MB. A translational vector from the zero DV to this identified corresponding position in the viewpoint reference frame is then formed and used as a DV predictor, as shown in Fig. 3. This predictor gives a better indication of where the optimal DVs are located, when compared to the median one adopted by the standard, thus allowing a reduction in the DE's search area and its associated computations. This has minimum impact on the coding efficiency and achieves a good gain in encoding speed. Since only median depth map values are utilized, this only indicates a good estimate of the optimal DV for the whole sub-MB partition, although now the actual RD optimal one should lie in the close neighborhood and still needs to be searched for through the DE process, but within a small search area. The optimally selected DVs are still transmitted as residual vectors from the median ones, to obtain H.264/MVC compatible bit-streams which can be decoded using the original decoder. Thus, this technique reduces the encoding computations while keeping a similar performance.

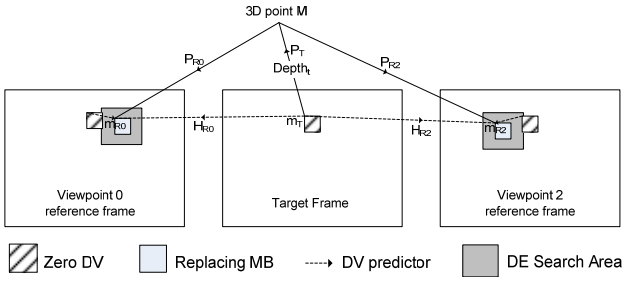

Fig. 3. Proposed optimal DV predictors, and their DE's search areas.

The picture on the left in Fig. 4 shows a frame from View 0, which is used as a viewpoint reference frame for the Target frame in View 2 presented on the right. This demonstrates that the proposed initial geometric DV marked a very accurate position from where to start searching for the optimal DV in the viewpoint reference frames. A preliminary exploration of the

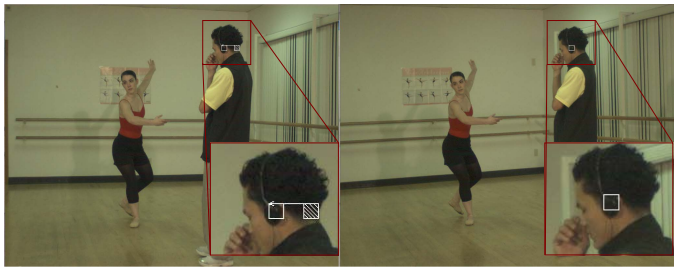

Fig. 4. MB indicated by the geometric disparity vector predictor.

Fig. 5. Inter-view mode separation due to depth change within the MB.

efficiency obtained by these corresponding areas identified by the multi-view equations, to form the geometric DV predictor, was reported in the authors' previous work [24], where this was reliable enough to sustain a reduced search area. The authors [25] showed also that this predictor is efficient to identify and project the encoded MVs from the encoded viewpoint reference frame to the target viewpoint. These estimates were used to form a new MV predictor that reduced the ME's computations. This DV predictor was also exploited by the authors [26] to identify the search area along the epipolar lines such that together they manage to reduce further the DE's search area. Furthermore, this predictor can be used to transmit the optimally selected DVs and obtain smaller residual vectors [27].

Nevertheless, along with this DV predictor, the potential optimal inter-view modes have to be identified and exploited to reduce further the DE's computations, by testing only the sub-optimal ones during RDO. This can be done by exploiting again the geometry available through the depth map data. This is because the partitioning of a MB that should be efficiently used to describe its compensation from a viewpoint reference frame also depends on the objects' depth and the camera setup. This occurs because, if two objects are close neighbors in the target viewpoint and are within the same depth level, they are expected to be found skewed but still exactly near each other in the viewpoint reference frames, as if they were the same object. On the other hand, if two objects are neighbors in the target viewpoint, but are within different depth levels, they are expected to be found far away, as shown in Fig. 5. Therefore, generally, for optimal disparity compensation, only the MBs that contain their partitions in different depth levels will use the partitioned modes to compensate their objects. So analyzing the change between the average depth map values of the MB's partitions, the potential optimal inter-view modes can also be predicted from the depth map MVV, and can be exploited to restrict the modes tested for RDO.

The proposed Inter-View Mode Selection (IVMS) process needs therefore to determine whether the basic partitions forming the modes are at different depth levels. This is used to determine if, and how to, partition the current MB, for an efficient inter-view compensation. Hence the sub-optimal DE's modes to test for RDO can be identified. Therefore, the target MB's area in its corresponding depth map frame is partitioned into its smallest main constituents possible, those of the 8×8 blocks, as shown Fig. 6(a). For every block, the depth map pels are averaged, divided by 10 and rounded to the nearest integer value, to identify properly the of the sub-MB's partitions at different depth levels. If all these depth averages are equal, it means that the MB is possibly representing an area with objects at almost a constant depth, or is representing the same object. Consequently, it is generally compensated as a whole MB from the viewpoint Reference frames, and the 16×16 compensation mode is used. If any of the two horizontal 8×8 sub-MBs are at the same depth level, it means that there is a possibility of an object with a different depth spanning through the horizontal. The 16×8 mode with its horizontal division can therefore help to efficiently compensate the two horizontally separated objects, and is tested as a possible optimal inter-view mode. If any of the two vertical 8×8 sub-MBs are equal, it means that the 8×16 mode with its vertical division is a possible inter-view mode. When neither both of the two horizontal sub-MBs nor both of the two vertical ones are equal, the 8×8 modes are tested too, since this area is usually represents a complex boundary. The majority of the MBs contains the background or are representing the same object, therefore, they are within the same depth level of the neighborhood, and thus, can be tested only with the 16×16 mode. However, for the objects' boundary that are at different depth or for a highly inclined areas, where there is evident depth change and a higher probability of having the compensated MB partitioned according to the objects' change in disparity, more than one mode division is tested.

Apart from these compensation modes, there are also the Intra and the SKIP modes that are less computational intensive but also very effective for DE. The 16×16 mode is the least intensive but still one of the most efficient R-D mode since it provides fair compensation distortion with the least transmission bits possible. Therefore, these three modes should always be tested first, and since they encode the whole MB, that one which minimizes the RDO cost is selected as the optimal one if the MB's 8×8 depths are constant. However, if a MB contains a depth change, the partitioned modes must also be tested according to this change, as discussed earlier. Sometimes, more than one partitioned mode has to be tested according to this change, to maintain the best possible R-D performance and coding efficiency. Thus, the flowchart of the inter-view mode selection process becomes as illustrated in Fig. 6(b). An adequate geometric DV predictor is calculated for every partition of the tested sub-optimal modes and this maintains further the coding efficiency, as each mode's partition is geometrically located in the viewpoint reference frame according to its local median correspondence.

Both the texture and the depth map MVV within a 3DV need to be encoded, and they contain similar MVV geometric properties [32]. Thus, the procedure to obtain an estimate of the optimal DV predictor and to select the sub-optimal modes can also be applied to speed-up the DE during depth map MVC, effectively obtaining similar speed-up. In this technique, only the averages of the 8×8 blocks are required from the depth map MVVs. As a result, even lower quality depth maps obtained after any compression algorithm that maintains the fidelity of the 8×8 average values (DC components), such as the H.264/MVC, or those produced by some depth map cameras, can be used. A smaller search area and multiple mode testing are still allowed for efficient encoding and to allow for possible depth inaccuracies in the depth map MVVs.

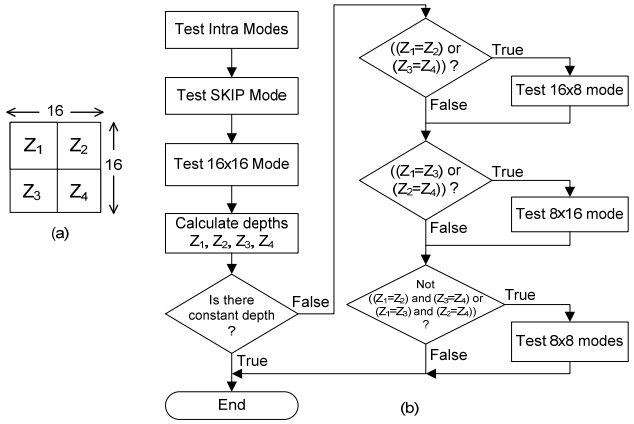

Fig. 6. (a) MB division for depth averaging, and (b) flow-chart of the IVMS process.

IV. Testing Methodology

To demonstrate the efficiency of the proposed geometric DE technique, it was implemented within the latest Joint Multi-view Video Coding model (JMVC v8.5) [33]. The encoder was modified to use either the multi-view or the Homogeneous equations, to identify the potential sub-MB's inter-view replacement and they were used to limit the optimal DV's search area. Furthermore, the encoder was updated to calculate the sub-MB partitions' average depth map values and they were used to limit the candidate DE's modes. The mode testing sequence was also altered as described above.

The 3DVs were encoded by the original and the modified encoders, and their speed-up gains were recorded. The original decoder was used to decode the bit-streams and a MVV evaluation was performed. The efficiency of the proposed sub-optimal DE technique was determined as a speed-up gain obtained when using a smaller search area with the geometric DV predictor, and finally as a further speed-up gain obtained when using the IVMS process as well, both over the original DE.

The *Breakdancers, Ballet, Balloons, Kendo* and *Newspaper* 3DVs were tested. The texture MVVs of the first two were captured using eight video cameras (YUV 4:2:0, 1024×768) arranged on an arc configuration and each view contains 100 frames, whilst for the others, capturing was done on a linear configuration and each view contains 300 frames. The provided depth map MVVs of the first two sequences were estimated using the method described by Zitnick *et al* [4], whilst for the others, the Depth Estimation Reference Software (DERS [34]) was used. For these simulations, the first three views from the first two 3DVs were compressed, while for the others, views three to five were considered. The MVC encoders were configured with the *Multi-view High profile*. Since the proposed fast DE is generally required for low-latency applications, the simulation parameters were chosen to obtain a low complexity encoder. To fairly analyze this fast DE with the IVMS process, only the DE was allowed. This was done to analyze correctly the loss in the inter-view coding efficiency obtained by the proposed geometric DE with the sub-optimal modes. The inter-view prediction structure was defined such that all frames in View 2/5 are forward predicted from View 0/3 and all frames in View 1/4 are bi-inter-view predicted from both View 0/3 and 2/5, to obtain the optimal inter-view coding efficiency [12]. The CAVLC was selected as the main entropy encoder to ensure further low delay characteristics [30]. For the original DE, a DV search range of ±32pels was used [35], while for the geometric DE, a smaller search range of ±10pels was chosen [24]. A quarter-pel accuracy estimation resolution was used. Both FSE and the diamond search as FASE [30] were used to determine the optimal DVs for both the texture and depth map MVC. Four Quantization Parameters (QPs); 28, 32, 36, and 40 were used to compare the R-D performances [35], [36]. All the simulations were carried out on a single core within a system powered by a 3.2 GHz central processing unit.

V. Experimental Results and Analyses

A. Rate-Distortion Influence vs. Speed-up Gain Analyses

The R-D performance curves plotted in Figs. 7 and 8, show the video quality as the average Peak Signal-to-Noise Ratio (*PSNR*), in dB**,** versus their average total MVC *bit-rate*, in kbps. These represent the average performance of the various tested techniques on the *Ballet* and the *Breakdancers* 3DVs, for their texture and the depth map MVC, respectively. These graphs compare the original encoder, the encoder with only the geometric DE (gDE) with its DV predictor calculated though the Homogenous equations, and the encoder with both the geometric DE and the IVMS process. The results also include the performance obtained when either the FSE or the FASE is used to acquire the optimal DVs. After analyzing these curves, one finds that they are very close to each other for both the texture and the depth map MVC, indicating that their RD performance is almost preserved by using the proposed techniques. A more detailed averaged performance is presented in Table I for the same 3DVs. This compares the average change in video quality, the percentage increase in the total MVV bit-rate, and the overall speed-up gain in the encoding time, compared to the original encoder that uses the FSE; since the latter combination offers the best possible RD performance. The overall speed-up gains represent the increase in time efficiency of the proposed fast DEs.

Although such RD curves and values are very popular to measure the efficiency of the video coding tools, the Bjøntegaard Delta [36] (BD) method is more suitable to determine this very small loss in PSNR or its equivalent bit-rate changes between the R-D curves. Thus, Table II presents a general summary of all the changes in R-D performances obtained by the final proposed technique, as BD-PSNR and BD-bit-rate values, between the original and the proposed methods, for all the tested 3DVs.

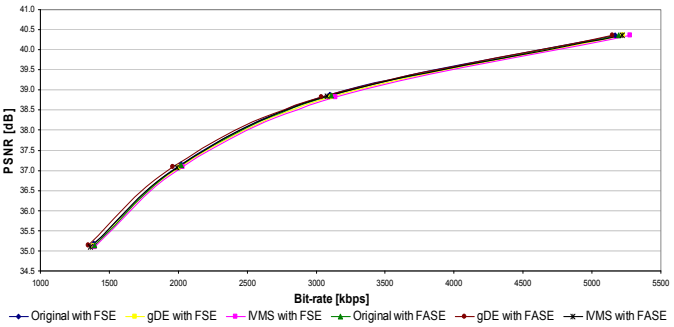

Fig. 7. Average R-D curves for the Texture MVVs of two tested 3DVs.

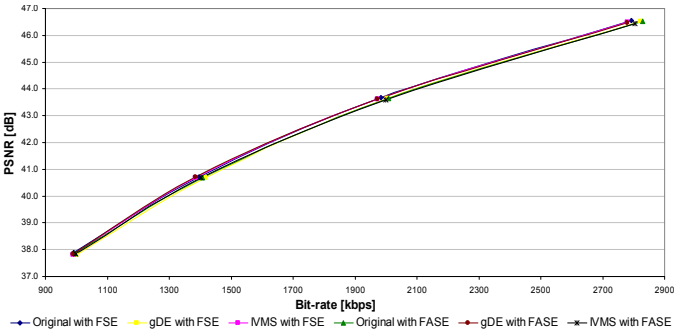

Fig. 8. Average R-D curves for the Depth map MVVs of two tested 3DVs.

From Table I, it can be deduced that after averaging the speed-up gains obtained for the texture and the depth map MVC of the two tested 3DVs, the decreased DE search area gives a significant average speed-up gain of about 7.4 times for FSE, and about 1.8 times for FASE, over the original DE. These gains were obtained because although there is a slight increase in computations, to calculate the new DV predictor, this is negligible compared to the resulting significant computational reduction obtained by a smaller search area; of about 10 times from the original number of search points and their calculations, during DE with FSE, and of about 4 times during DE with FASE. Furthermore, using the IVMS process provides a further speed-up gain of 3.6 for FSE, and of 2.4 for FASE, over the geometric DE, respectively. These gains were obtained by another significant DE's computational reduction of about 4 times when fewer modes were tested during RDO; with the reduced statistics indicated in Table III. Table II shows that a final average overall speed-up gain of about 26.0 times during FSE, and of about 4.5 times during FASE, over the original DE can be achieved. These gains were obtained with only an average MVV quality loss of about 0.08 dB and 0.17 dB BD-PSNR during texture and depth map MVC, respectively. Out of these, a degradation of up to 0.05 dB and 0.12 dB, respectively resulted from the geometric DE alone. This represents an equivalent overall MVC BD-bit-rate increase of 1.8 % in the texture's bit-rates and 2.1 % in the depth map's bit-rates. Thus, the original average DE efficiencies, with respect to Intra coding; of about 34 % reduction in high bit-rates and 52 % in low bit-rates for texture MVC, and of about 40 % in high bit-rates and 62 % in low bit-rates for depth map MVC, were almost preserved for the inter-view predicted viewpoints. Therefore, the proposed search area and the inter-view mode selection process demonstrated efficient to limit the candidate search points and modes tested for RDO. Furthermore, these results show that this technique can be used with fast search estimation strategies, such as the diamond search, to obtain larger speed-up gains. Again the gains were achieved without any significant influence on the coding efficiency. These methods resulted efficient for both the texture and the depth map MVC, and can be effectively applied to encode both MVV types within the 3DVs.

TABLE II
SUMMARY OF THE FINAL RESULTS OBTAINED FOR THE EVALUATED 3DVS

3DV Sequences	DB-PSNR (dB)	DB-Bit-rate (kbps)	Overall FSE Speed-up (×)	Overall FASE Speed-up (×)
Texture MMV of the:				
Ballet	- 0.093	+ 1.95 %	24.68	4.07
Breakdancers	- 0.064	+ 1.45 %	25.19	4.15
Balloons	- 0.084	+ 2.13 %	24.01	4.25
Kendo	- 0.099	+ 1.90 %	23.59	4.37
Newspaper	- 0.056	+ 1.64 %	22.54	4.68
Depth map MVV of the:				
Ballet	- 0.160	+ 2.57 %	28.70	4.85
Breakdancers	- 0.148	+ 2.01 %	28.47	4.08
Balloons	- 0.197	+ 2.10 %	27.86	4.71
Kendo	- 0.183	+ 2.23 %	27.54	4.62
Newspaper	- 0.156	+ 1.94 %	27.23	4.31

TABLE I
A DETAILED SUMMARY OF THE AVERAGE LOSSES IN CODING PERFORMANCE AND SPEED-UP GAINS FOR TWO TESTED 3DVS

Sequences	Original with FSE	Change in	gDE with FSE	IVMS with FSE	Original with FASE	gDE with FASE	IVMS with FASE
Ballet	38.31 dB	PSNR (dB)	- 0.026	- 0.046	- 0.020	- 0.052	- 0.056
Texture	3048.50 kbps	Bit-rate (%)	+ 0.34	+ 1.74	+ 0.48	+ 1.33	+ 2.20
MVV	16.56 hrs	Speed-up (×)	8.09	24.68	10.49	21.77	49.30
Ballet	41.84 dB	PSNR (dB)	- 0.046	- 0.049	- 0.024	- 0.073	- 0.086
Depth Map	2129.40 kbps	Bit-rate (%)	+ 1.28	+ 2.09	+ 1.03	+ 2.11	+ 2.12
MVV	17.58 hrs	Speed-up (×)	7.15	28.70	13.49	26.31	65.42
Breakdancers	37.47 dB	PSNR (dB)	- 0.032	- 0.035	- 0.013	- 0.029	- 0.045
Texture	2789.72 kbps	Bit-rate (%)	+ 0.15	+ 1.33	+ 0.06	+ 0.82	+ 2.09
MVV	18.37 hrs	Speed-up (×)	7.18	25.19	13.41	23.33	56.84
Breakdancers	42.57 dB	PSNR (dB)	- 0.032	- 0.064	- 0.027	- 0.026	- 0.036
Depth Map	1452.26 kbps	Bit-rate (%)	+ 0.83	+ 1.68	+ 0.85	+ 1.35	+ 1.43
MVV	17.61 hrs	Speed-up (×)	6.98	28.47	14.67	25.51	60.82

B. Computational Cost Analysis

The proposed technique imposes only a slight computational burden on the MVC encoder. The average MB's computations involved to determine the average depth map value for every 8×8 partition, and to accordingly determine the sub-optimal modes to test to disparity estimate the target MB, are listed in Table III (a). Those involved to determine the corresponding areas through the Homogenous or the Multi-view equations are listed in column (b) and (c), respectively. To perform this method, the system requires only the corresponding depth map frame of the target viewpoint, thus, it only requires loading an extra frame in memory. However, when the reduced search area and the sub-optimal modes are used, they eliminate a proportionally huge amount of search points to test during RDO, where each test point requires the average computations presented in Table IV. When a mode is not tested for RDO, all of its computations; consisting mainly of its constituent sub-MBs' estimation in every test point within the search area and in every reference frame, are completely eliminated. Thus, the computational reduction becomes proportional to the reduction in the search area and the restriction of the modes, which both contribute significantly to the final MVC speed-up gains.

TABLE III
COMPUTATIONS REQUIRED FOR THE GEOMETRY IN THE PROPOSED METHOD

	(a) 8×8 averages	(b) Multi-view Proj.	(c) Homo. Proj.
Additions (+/-)	784	32	4
Multiplications (*)	0	34	6
Divisions (/)	4	5	2

TABLE IV
COMPUTATIONS REQUIRED FOR EVERY TESTED SEARCH POINT

Additions (+/-)	896
Multiplications (*)	11
Divisions (/)	16

The average computations involved in performing a complete search for the optimal RD data, to disparity estimate the target MB in an I-B-P inter-view coding structure, are summarized in Table V. This includes the computational burden of the original method, that of the geometric DE which calculates the corresponding areas through the multi-view geometry (1), or the Homogeneous equations (2), and that of the latter geometric DE together with the IVMS process. This table demonstrates that the proposed geometric DE reduces to about one tenth, the highest order computations of the RDO process. These are reduced by another factor of 4, when the sub-optimal inter-view mode selection process is used. All these reductions are independent from the power of the device's processing unit, making the solution compatible to any consumer device. However, lower/higher clocked devices are then expected to take longer/shorter encoding and decoding durations than those presented. Nevertheless, comparable processing capabilities are incorporated in today's consumer devices, as seen with; latest mobile phones/tablets which typically incorporate a dual core 1.4 GHz processor, latest embedded processors reaching 2.5 GHz speeds, and some consumer electronics being powered by a 3.2 GHz low power processor, especially those used for video applications.

Using these sub-optimal techniques will neither affect the decoder's complexity nor its computations, so the original decoder can be used. However, if these geometric DV predictors are used to efficiently encode the transmitted optimal vectors, to obtain smaller residual disparity vectors and thus reduce the encoded MVC bit-rates; as demonstrated by the same authors in [23], the predictors will have to be re-calculated at the decoder, and this will influence its complexity. The last two rows of Table V demonstrate this impact. It shows that while the encoder is slightly affected by the choice between the multi-view and the Homogeneous equations, the decoder is highly influenced. In fact, the decoder that uses the multi-view equations had an increase of about 11 % in its frame decoding duration, which then goes down to only 2.5 % when utilizing the Homogenous equations. Using either method will not influence the projection values or the DE's coding efficiency, as the latter equations are calculated for every integral depth map value and their used depth map average values are represented by integers.

Table VI gives the probability statistics of each mode being tested and selected during the RDO process, by either the original or the proposed methods. These statistics were collected over the five tested 3DVs and were also used to determine the average computation required per MB, presented in Table V. This table demonstrates that although the modes are tested less frequently, the modes' selection percentages obtained by the sub-optimal IVMS process are very close to the original ones.

TABLE V
AVERAGE AMOUNT OF THE COMPUTATIONS REQUIRED FOR THE RATE-DISTORTION OPTIMIZATION OF AN INTER-VIEW PREDICTED MB

Method	Computations	Original	geometric DE with Multi-view Eqns.	geometric DE with Homogeneous Eqns.	geometric DE with Homogeneous equations and IVMS
DV Predictors	Additions (+/-)	N/A	1647	1322	41
	Multiplications (*)		459	35	28
	Divisions (/)		68	62	19
Mode estimation	Additions (+/-)	N/A	N/A	N/A	784
	Multiplications (*)				0
	Divisions (/)				4
R-D optimization Search Area points	Table IV	32pel×32pel×8modes ×1.5 (ref)	10pel×10pel×8modes ×1.5 (ref)	10pel×10pel×8modes ×1.5 (ref)	10pel×10pel×1.85modes ×1.5 (ref)
R-D optimization Computations	Encoding $O(n^2)$	331,776	32,927	32,481	7,569
	Decoding $O(n^2)$ [47]	0	203 †	47 †	46 †
	Decoding duration (msec)* [47]	156	173	160	159

* Decoding duration of a whole frame with 3072 MBs.

† Calculations estimated based on the mode statistics presented in Table VI.

TABLE VI
MODE SELECTION BY THE ORIGINAL AND THE PROPOSED METHODS

Probabilities of mode selection	16×16	16×8	8×16	8×8
Mode tested for RDO - Original	100%	100%	100%	100%
Mode is tested for RDO - Proposed	100%	17%	16%	13%
Mode selected as optimal - Original	67%	13%	11%	9%
Mode selected as optimal - Proposed	71%	11%	10%	8%

Fig. 9. MB partitions depending on the 8×8 changes in the depth map.

Fig. 10. R-D optimally selected modes for the texture data frame.

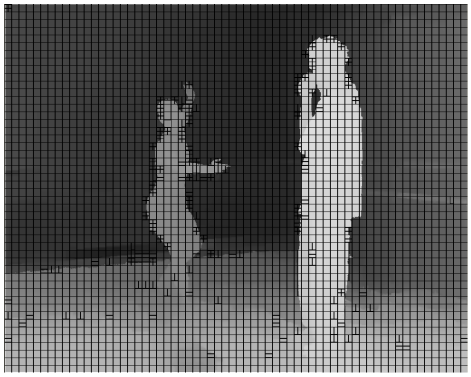

Fig. 11. R-D optimally selected modes for the depth map frame.

C. Mode Division Analyses

Fig. 9 shows the first Texture frame of View 2 from the *Ballet* 3DV, and this is marked with a white line where the corresponding 8×8 sub-MB's partitions contain separate depth levels, such that the potential modes for that MB can be visualized. This illustrates that the more complex modes are tested around the objects' boundaries and across evident/ abrupt depth map changes where the MBs should be partitioned for efficient inter-view compensation. The division of the MBs follows exactly the objects' boundary, so the tested modes should give efficient compensation. Fig. 10 and Fig. 11 show the R-D optimally selected modes through the original DE of frame 1 of the texture and the depth map MVC, respectively. These two images illustrate the modes selected when the Intra coding and the SKIP modes were disabled, and only disparity compensation was allowed. A QP of 30 was used, where a texture quality of around 40.2 dB and a depth map quality of 44.5 dB were obtained, which are among those that require the most complex modes' division and achieve good compression quality. This was performed to study the actual optimal compensation modes that the DE process would select as optimal to compensate the MBs in the frames. These confirm that the majority of the compensated MBs' divisions do in fact occur around the objects' boundaries or across evident/abrupt depth changes, and that the majority of these modes are predictable. In fact, on average, about 91 % of the partitioned modes obtained by the texture MVC, and 95 % of those obtained by the depth map MVC, are correctly predicted as the R-D optimally selected ones, when these are averaged over all the tested 3DVs and their tested QPs. The small influence on the coding efficiency demonstrates that the modes that are not correctly predicted, are also efficiently encoded by other modes, however, as these are not optimal, they cause the small registered loss in coding efficiency.

VI. CONCLUSION

This paper presented a fast disparity estimation method that exploited depth map data, already available in the transmitted 3D videos, to obtain a better disparity vector predictor and a sub-optimal inter-view mode selection process. These were then used to significantly improve the disparity estimation's speed. Simulation results confirm that this technique is efficient enough and reduces about 95 % of the computational time required for the original encoding with negligible influence on the rate-distortion performance. Around 68 % of this reduction came from the geometric prediction while about 32 % of it came from the sub-optimal mode decision process. These computational reductions allow more utilization of 3D video coding systems and their services in everyday-to-day applications and on today's consumer's devices.

ACKNOWLEDGMENT

The authors would like to thank the Microsoft Research (MSR) group, the Nagoya University and the Gwangju Institute of Science and Technology (GIST) for providing the Multi-view Video plus Depth 3D video test sequences.

REFERENCES

[1] L. Onural, "Television in 3-D: What are the prospects?," *Proc. of IEEE*, vol. 95, no. 6, pp. 1143-1145, Jun. 2007.

[2] M. Tanimoto, "Overview of Free Viewpoint Television," *Signal Process.: Image Comm.*, vol. 21, no. 6, pp. 454-461, Jul. 2006.

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

[4] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winderm, and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM SIGGRAPH & ACM trans. on Graph.*, pp. 600-608, Aug. 2004.

[5] C. Weigel, and N. Treutner, "Flexible openCL accelerated disparity estimation for video communication applications," *in Proc. of 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*, Antalya, Turkey, May 2011.
[6] 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 Adv. in Signal Process.*, vol. 2009, 13 pages.
[7] C. Hellge, E. G. Torre, D. G.-Barquero, T. Schierl, and T. Wiegand, "HDTV and 3DVT services over DVB-T2 using multiple PLPs with SVC and MVC," *in Annual IEEE Broadcast Symposium*, 2011.
[8] 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.
[9] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient prediction structures for multi-view video coding," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 17, no. 11, pp. 1461-1473, Nov. 2007.
[10] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," *in Proc. of IEEE International Conference on Image Processing*, Texas, USA, pp. 201-204, Sept. 2007.
[11] R. Pan, Z.-X. Hou, and Y. Liu, "Fast algorithms for inter-view prediction of multiview video coding," *Journal of Multimedia*, vol. 6, no. 2, pp. 191-201, Apr. 2011.
[12] ISO/IEC MPEG, and ITU-T VCEG, "*Survey of algorithms used for Multi-view Video Coding (MVC)*," Doc. N6909, Jan. 2005.
[13] X. San, H. Cai, J. -G. Lou, and J. Li, "Multiview image coding based on geometric prediction," *IEEE Trans. Circuits Syst. Video Techno.*, vol. 17, no. 11, pp. 1536-1548, Nov. 2007.
[14] Y. T. Kim, J. Y. Kim, and K. H. Sohn, "Fast disparity and motion estimation for multiview video coding," *IEEE Trans. Consumer Electron.*, vol. 53, no. 2, pp. 712-719, May 2007.
[15] X. Li, D. Zhao, X. Li, Q. Wang, and W. Gao, "Fast disparity and motion estimation based on correlations for multi-view video coding," *IEEE Trans. Consumer Electron.*, vol. 54, no. 4, pp. 2037-2044, Nov. 2008.
[16] W. Zhu, X. Tian, F. Zhou, and Y. Chen, "Fast disparity estimation using spatio-temporal correlation of disparity field for multiview video coding," *IEEE Trans. Consumer Electron.*, vol. 56, no. 2, pp. 957-964, May 2010.
[17] Z. J. Peng, G. Y. Jiang, and M. Yu, "A fast multiview video coding algorithm based on dynamic multi-threshold," *in Proc. of International Conference on Multimedia and Expo*, New York, USA, pp. 113-116, Jun. 2009.
[18] M. Yu, Z. Peng, W. Liu, F. Shao, G. Jiag, and Y. D. Kim, "Fast macroblock selection algorithm for multiview video coding based on inter-view global disparity," *in Proc. of Congress on Image and Signal Procressing*, Sanya, Hainan, pp. 575-578, May 2008.
[19] D. –H. Han, and Y. –L. Lee, "Fast mode decision using disparity vector for multiview video coding," *in Proc. of International Conference on Future Generation Communication and Networking Symposium*, vol. 3, pp. 209-213, Dec. 2008.
[20] L. Shen, Z. Liu, T. Yan, Z. Zhang, and P. An, "View-adaptive motion estimation and disparity estimation for low complexity multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 925-930, Jun. 2010.
[21] L. Shen, Z. Liu, P. An, R. Ma, and Z. Zhang, "Low-complexity mode decision for MVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 6, pp. 837-843, Jun. 2011.
[22] G. Yang, L. Liang, and W. Gao, "An epipolar restricted inter-mode selection for stereoscopic video encoding," *in Proc. of Picture Coding Symposium,* Nagoya, Japan, pp. 338-341, Dec. 2010.
[23] B. W. Micallef, C. J. Debono, and R. A. Farrugia, "Fast Inter-Mode Decision in Multi-view Video plus Depth Coding," *in Proc. of Picture Coding Symposium*, Kraków, Poland, pp. 113-116, May 2012.
[24] B. W. Micallef, C. J. Debono, and R. A. Farrugia, "Exploiting depth information for fast multi-view video coding," *in Proc. of Picture Coding Symposium*, Nagoya, Japan, pp. 38-41, Dec. 2010.
[25] 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: The True Vision - Capture, Transmission and Display of 3D Video*, Antalya, Turkey, May 2011.
[26] B. W. Micallef, C. J. Debono, and R. A. Farrugia, "Fast disparity estimation for multi-view video plus depth coding," *in Proc. of International Conference on Visual Communication and Image Processing*, Tainan, Taiwan, Nov. 2011.
[27] B. W. Micallef, C. J. Debono, and R. A. Farrugia, "Exploiting depth information for efficient multi-view video coding," *in Proc. of International Conference on Multimedia and Expo*, Barcelona, Spain, Jul. 2011.
[28] ISO/IEC IS 14496-10, *Advanced Video Coding for Generic Audiovisual Services*, ITU-T Rec. H.264, Mar. 2009.
[29] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 688-703, Jul. 2003.
[30] I. E. G. Richardson, *H.264 and MPEG-4 Video Compression, Video Coding for Next-generation Multimedia*, John Wiley & Sons, 2003, UK.
[31] 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.
[32] R. Hartley, and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2003, UK, pp. 279-309.
[33] ISO/IEC MPEG, and ITU-T VCEG, "*Joint multi-view video coding model (JMVC 8.5)*", Mar. 2012.
[34] ISO/IEC MPEG, "*3DV depth estimation and view synthesis software package*," Doc. N12188, Jul. 2011.
[35] ISO/IEC MPEG, and ITU-T VCEG, "*Common Test Conditions for Multiview Video Coding*," Doc. JVT-U211, Oct. 2006.
[36] G. Bjøntegaard, "*Calculation of average PSNR differences between RD-curves,*" Doc. VCEG-M33, Apr. 2001.

BIOGRAPHIES

Brian W. Micallef (S'08, M'13) received the first degree in Electrical Engineering from the University of Malta, Malta, in 2009. He is currently pursuing his Ph.D. degree from the Faculty of Information and Communications Technology at the University of Malta.

His research interests are multi-view video coding, resilient multimedia transmission, multimedia networking and image processing.

Carl James Debono (S'97, M'01, SM'07) received his B.Eng. (Hons.) degree in Electrical Engineering from the University of Malta, Malta, in 1997 and the Ph.D. degree in Electronics and Computer Engineering from the University of Pavia, Italy, in 2000.

Between 1997 and 2001 he was employed as a Research Engineer in the area of Integrated Circuit Design with the Department of Microelectronics at the University of Malta. In 2000 he was also engaged as a Research Associate with Texas A&M University, Texas. In 2001 he was appointed Lecturer with the Department of Communications and Computer Engineering at the University of Malta and is now an Associate Professor. He is currently the Deputy Dean of the Faculty of ICT at the University of Malta.

Prof. Debono is a senior member of the IEEE and served as chair of the IEEE Malta Section between 2007 and 2010. He is the IEEE Region 8 Vice-Chair of Technical Activities for 2013. He is also a member of the management committee of the COST Action IC1105 - 3D Content Creation, Coding and Transmission over Future Media Networks (3D-ConTourNet) where he chairs the 3D Media Coding Working Group. His research interests are in wireless systems design and applications, multi-view video coding, resilient multimedia transmission and modeling of communication systems.

Reuben A. Farrugia (S'04, M'09) received the first degree in Electrical Engineering from the University of Malta, Malta, in 2004, and the Ph.D. degree from the University of Malta, Malta, in 2009.

In 2004 he was employed as a Research Engineer with the Department of Communications and Computer Engineering of the University of Malta. His research work included wireless network modeling and resilient video coding. In January 2008 he was appointed Assistant Lecturer with the same department and is now a Lecturer. His research interests are in resilient video coding, distributed video coding and image processing.