COST Action IC1105: 3D-ConTourNet

3D Content Creation, Coding and Transmission over Future

3D Media Coding:
State-of-the-art and Research Directions

White Paper

Prepared by Working Group 2 – 3D Media Coding

Contents

WG2 – 3D Media Coding

White Paper – Version 2

Introduction

The last few years, have seen several advances in technology that have made the capture, transmission and display of 3D content feasible. Until recently 3D content was only available in specialized cinemas due to the investment needed and the bandwidth requirements. Current technology and infrastructure allows for the transmission of stereoscopic video over satellite, Blu-Ray™ disks, and Internet technologies [1]. These are target to stereoscopic displays where the user needs special glasses to present the right content to the human vision system. However, to provide a more immersive service and experience, more content coming from multiple cameras needs to be sent to the new auto-stereoscopic displays. Similarly, audio capturing needs more feeds to provide the 3D audio experience and ideally reproduce the original wave field. Furthermore, depth data is required to provide higher fidelity 3D TeleVision (3D TV) and allow for Free-viewpoint applications, where the end user can select the 3D view out of many [2]. These demand huge amounts of data that need to be delivered in minimum time over bandwidth-limited or space-limited channels. Therefore, 3D Media Coding plays an important role in making immersive 3D communications possible.

Standardization

Frame compatible format and simulcast coding of individual views involve a single view video encoder, such as H.264/AVC. This encoder relies on reducing the redundancies present both in space and in time between the current frame being encoded and its previous frame. The result of these schemes are limited, given that the former scheme results in loss of resolution and the latter is coding inefficient.

This has led the Moving Pictures Expert Group (MPEG) to provide better coding structures to improve compression efficiency. Since multiple cameras are capturing the same scene from distinct locations, these videos have high correlation also in between views. This can be exploited in a similar way to temporal coding in single view, this time using disparity compensation. This resulted in the standardization of the *Multiview High Profile* and the *Stereo High Profile*, in Annex H of the H.264/AVC standard. This Multi-view Video Coding (MVC) standard achieves around 30% bit rate gain in coding of the auxiliary view [3] compared to simulcast. The frame structure for three views, with their dependencies, is shown below for a Group Of Pictures (GOP) of 8.

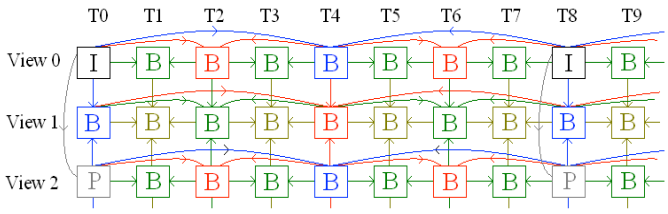

Figure 1 – Multi-view video coding hierarchical bi-prediction structure

This work is also evident in the recently approved video coding standard for High Efficiency Video Coding (HEVC) [4] where Annex F deals with multi-view video coding (MV-HEVC). This results in around 30% ~ 50% bit rate savings compared to previous H.264 products.

Data Representation

Stereoscopic 3DTV Representation

The stereoscopic representation is the currently most popular solution to broadcast 3D videos. This scheme incorporates the transmission of the left eye view and the right eye view sequentially. These are sent either side-by-side or top-and-bottom in the same encapsulation space of high definition television, using frame compatible formats. Stereoscopic displays are then used for display.

Model-Based Representation

Model-based approaches classify generic models that contain segments which can be represented using closed meshes, for example triangular meshes [5]. These models require adaptation to better represent the object in the scene. This is done through scaling of individual segments together with some deformation of the surfaces.

This method maps the input streams into texture space, effectively transforming the 3D model into a 2D representation. Cuts on the surface are needed for general meshes to limit distortion, resulting in distinct patches [5]. The texture maps coming from each camera view are then encoded using techniques like 4D-SPIHT encoding [5, 6]. The output is thus a set of camera views encoded at every time step.

Point Sample-Based Representation

Point sample methods map 2D video processing to 3D video polygonal representation. An example of such a scheme is found in [7], where a differential update technique makes use of the spatio-temporal coherence of video scenes captured by the multiple cameras. Operators on 3D video fragments are used to compensate for changes in the input. These operators are later merged with the video stream for transmission [7]. A surface normal vector in spherical coordinates together with a color value is used to define a 3D video fragment. The camera parameters and identifiers, and image coordinates of the 2D pixel are required to allow geometrical reconstruction and thus have to be transmitted with the video. Entropy encoding of the vectors and color is then applied to achieve further compression of the data [7].

Multi-view Video Representation

Multiple videos capturing the same scene from different angles constitute the Multi-View Video (MVV) representation. This type of data can be coded using Multi-view Video Coding, which defines the way to exploit spatial, temporal, and inter-view redundancies

present in viewpoint videos captured at the same instant from multiple cameras for compression. MVC has been included as an extension to the H.264/AVC and HEVC standards [8, 9], where the motion estimation process is extended to estimate also a macroblock in a frame from another viewpoint. This allows for the selection of the better estimation vector between motion and disparity [8].

The best macroblock replacement is found using the Langrangian Rate-Distortion (RD) cost function [10] on the candidate list. Thus, encoding of a macroblock is done by first identifying the best sub-macroblock estimates, and if available, encode their translational vectors and the block's difference as residual data. If there is no good estimate, then INTRA coding is selected. The more compensated macroblocks within a frame, the more efficient the encoding is. However, these exhaustive searches for motion, disparity and mode compensation are highly computational intensive which implies that non-optimal solutions are needed to keep encoding within real-time parameters [11-14]. The coding architecture uses either the Low Latency or the Hierarchical Bi-prediction structures to allow for different applications. The former is ideal for transmission of time constraint data, such as videoconferencing, while the latter allows for better coding efficiency and can be used in buffered applications, such as video streaming. Entropy coding is then applied to further reduce redundancy, where the available schemes are: Context Adaptive Variable Length Coding (CAVLC) or Context Adaptive Binary Arithmetic Coding (CABAC).

Multi-view Video plus Depth Representation

The Multi-View Video plus Depth (MVD) format allows the use of depth maps in conjunction with the texture video. This provides geometrical information that can be used for better coding and view reconstruction, which provides support to wide angle and auto-stereoscopic displays [15]. This format is analogous to the MVC system but in this case it also includes the geometric per-pixel depth maps of the texture incorporated within the data stream. The data within the depth video permits the utilization of Depth Image-Based Rendering (DIBR) techniques for the reconstruction of viewpoints in between the coded ones, thus allowing free-viewpoint video applications. This format [13] is expected to be the main format for transmission of 3D Videos (3DVs) in the 3D extension of HEVC standard [4].

Coding Schemes

Pattern Matching-Based Algorithms

A simple way to encode images is by using pattern matching and representation from a dictionary. The dictionary can be increased and adapted to the image patterns that need representation. Scale transformations are used to allow exploitation of similarities within the images. The concept is shown in Figure 2.

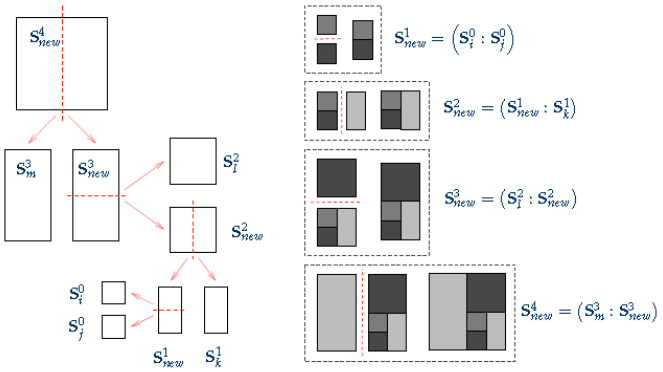

Figure 2 – Dictionary search

To reduce the search time, the dictionary search algorithm can be preceded by a prediction step as shown in Figure 3. The method can be applied to stereo images, depth map data, and multi-view video.

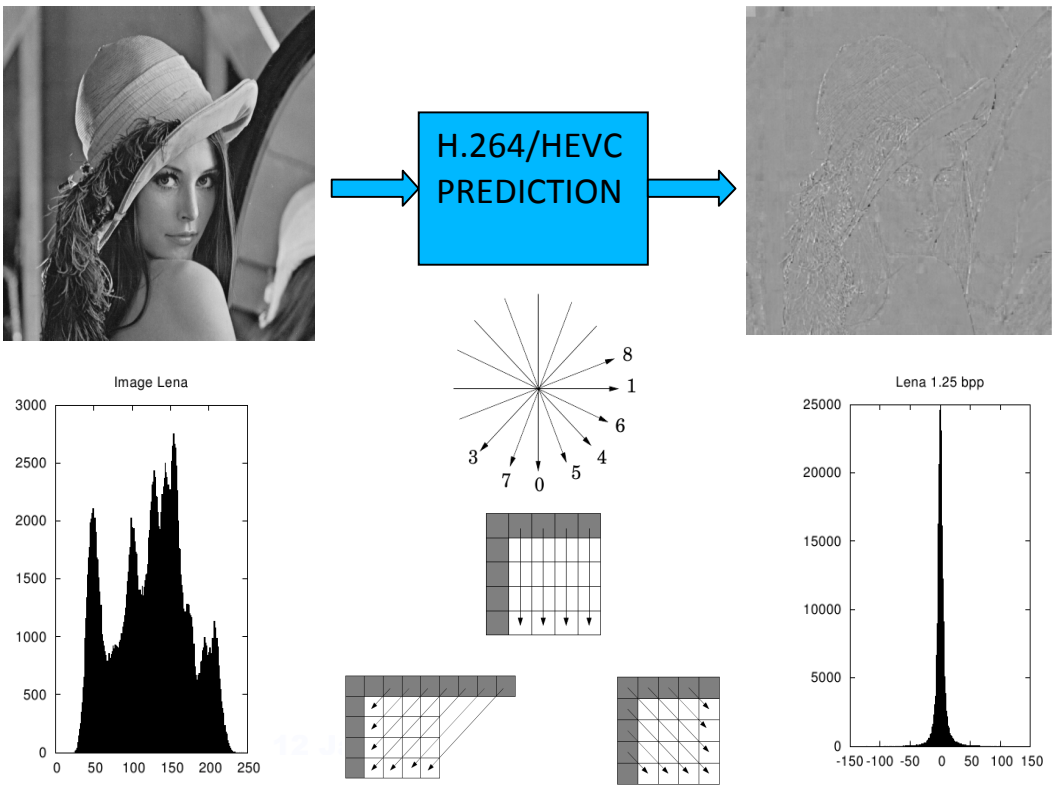

Figure 3 – Application of prediction techniques before dictionary encoding

Multi-view Video plus Depth Coding

Multi-view texture and depth map videos are needed to provide fidelity in 3D TV and allow Free-viewpoint services. This allows for arbitrary view synthesis in between the transmitted camera views, through DIBR. The process is illustrated in Figure 4, where using the depth map data and the multi-view geometry, pixels are correctly located in the

required viewpoint for its synthesis. The texture and the depth video can both be encoded using the multi-view video coding structure. However, the depth data requires more protection of the edges as its function is mainly for DIBR whose performance depends on the correct information of the edges.

Figure 4 – Illustration of view synthesis

Depth Map Compression by "Don't Care" Regions

In view rendering, what really matters is the quality of the synthesized view and not the depth map. This means that a range of error on the depth data can be tolerated within certain regions. The region of tolerance is defined as the "don't care" region, as shown in Figure 5. The coding is then done by transmitting the difference prediction error with respect to the "don't care" region instead of the actual depth pixel value. Coding statistics and visual representation are shown in Figure 6.

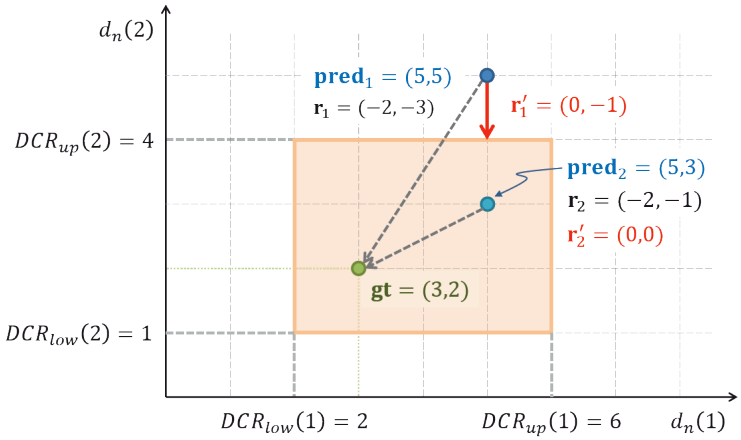

Figure 5 – Representation of the "don't care" region

Table 1. Coding statistics for two RD points of *Kendo*

	bitrate [kbps]	PSNR [dB]	% SKIP	Motion info. [bit/frame]	Residuals [bit/frame]
no DCR	230.4	33.99	80.20	582.10	522.41
DCR ($\tau = 5$)	179.5	34.04	92.25	253.48	240.62

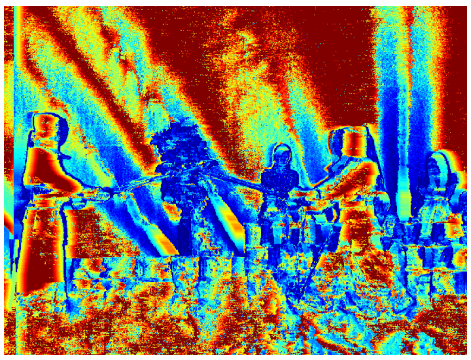

Figure 6 – Coding statistics and visual representation

Depth Map Compression by Exploiting Depth Inheritance

The HEVC Most Probable Mode (MPM) can be used for INTRA depth coding. The list of modes can be modified to include the mode of the co-located texture block. This is effective when the texture and depth blocks have a contour of a single dominant direction. The histogram of the module and angle of the image gradient can be used to detect blocks having a single dominant direction. Two criterions can be used, namely: the SOBELMAX and the SOBELHIST criterions. Sobel filtering on the texture block that corresponds to the current depth block is performed first. This gives two matrices G_x and G_y which are used to calculate the module and the angles matrices. The SOBELMAX criterion is the maximum value in a module matrix. It gives a measure of the amplitude of the edges present in the texture block but does not provide knowledge if those edges are directional or not. The SOBELHIST criterion considers the angles of these edges giving the directionality. The SOBELHIST is found by establishing the histogram of the angles in the angle matrix, and only select the angles with a corresponding module which is greater than threshold, to eliminate noisy components. If there is only one clear edge in the block, there is only one peak in the histogram and the criterion should return a high value. If there are many directional edges, there are many peaks returning a low value. Results are shown in Figure 7.

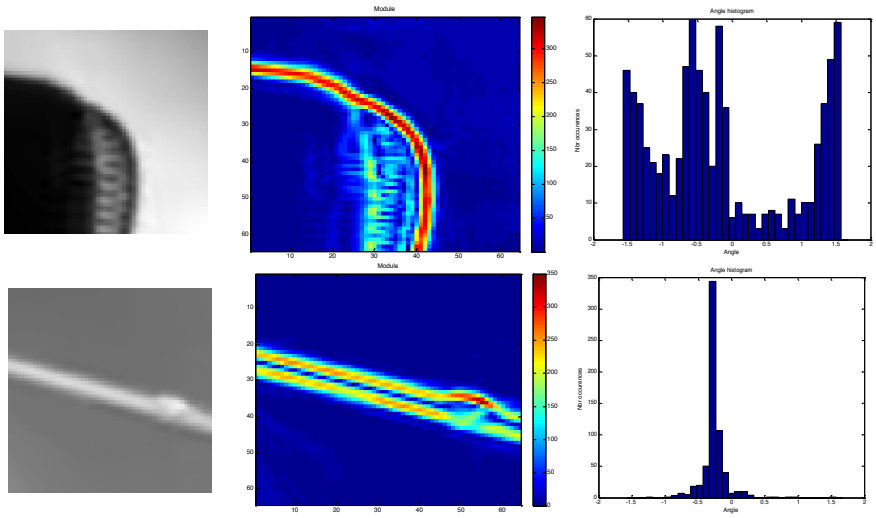

Figure 7 – Single dominant edge

Depth Map Coding in 3D-HEVC

The depth map coding can apply inter-view disparity compensation prediction similar to the texture coding. Depth maps can also inherit motion vectors from the co-located textures. This means that it is possible to have inter-component prediction between texture and depth map videos. An example of the prediction of block partitions based on the co-located texture images is shown in Figure 8 [17].

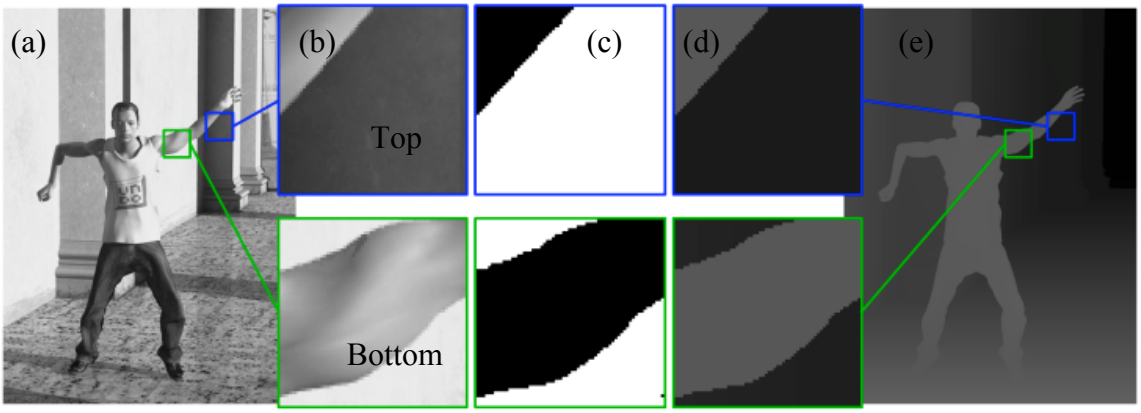

Figure 8 – Wedgelet (Top), Contour (bottom), (a) texture, (b) coding unit for texture, (c) Mean absolute deviation of (b), (d) coding unit for depth map, (e) depth map.

The focus of the rate distortion optimization needs to be focused on the synthesized views and not the depth maps themselves. This is because the depth maps are not viewed by the user and therefore the resulting synthesized view has to be considered [18]. The distortion change can be found as described in Figure 9.

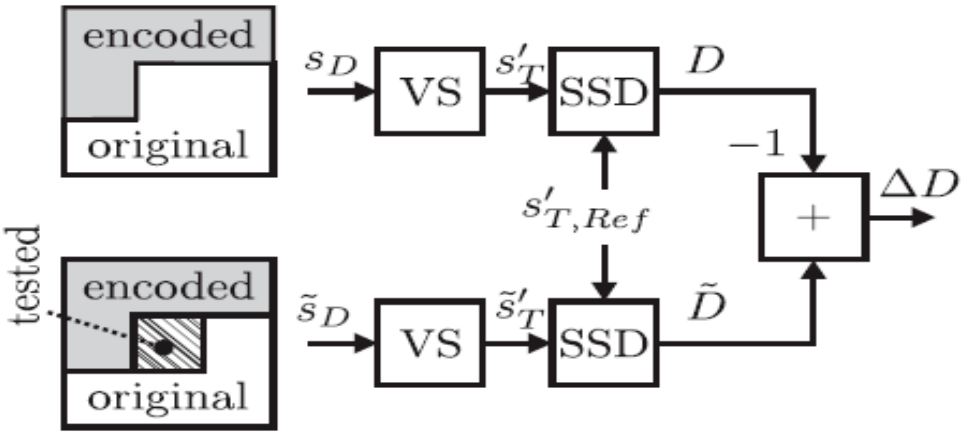

Figure 9 - Synthesized view distortion change related to the distorted depth data of the tested block, VS stands for View Synthesis.

3D Holoscopic Imaging

A 3D Holoscopic image is made up of a large number of closely packed distinct micro-images. These images are then viewed through an array of spherical convex lenses, as shown in Figure 10. This technology can be used also in 3D Holoscopic videos. The coding efficiency can be improved by exploiting the inherent cross-correlation of the 3D holoscopic images. This means that within the current frame, a similar reference picture can be found. Block-based matching can be applied as shown in Figure 11, to determine a predictor with the strongest similarity within the current macroblock partition.

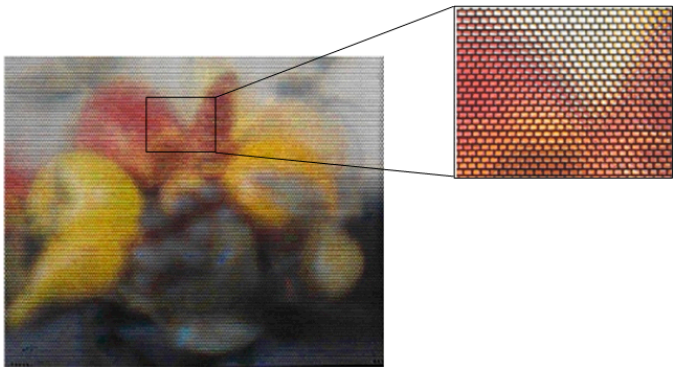

Figure 10 – 3D Holoscopic image

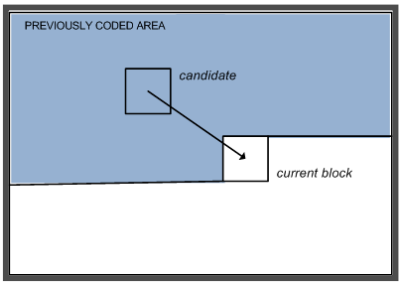

Figure 11 – Prediction vector

Research Trends

Although a lot of research effort has been directed towards 3D video coding, there is still a lot of work needed to increase the data being transmitted over bandwidth-limited channels and allow processing on limited-processing devices. Furthermore, more effort is needed towards improving the quality of experience of the user. This section highlights current trends in research in this field.

Image and Video Coding

Better pattern matching algorithms can be applied to improve the 3D image and video coding efficiency. This can be coupled with efficient predictive coding. Furthermore, the geometric transforms can be developed and applied to motion and disparity compensation of 3D content.

Depth Map Coding

More study is required for depth map coding. Initial studies had assumed that depth information is not so important and could be compressed heavily. However, later it was found that edges need to be preserved to ensure adequate quality of the synthesized views. Thus, the structure of the depth maps needs to be considered by using different coding quality for the edges and the smooth regions. A scalable approach can be a solution to this problem. Furthermore, down/up-scaling of the depth maps can be used to reduce the transmitted data. This will require joint up-scaling by using the texture information.

Texture plus Depth Video Coding

Coding techniques have to consider the impact of contours in the depth data. Inter-component correlation can be better exploited for compression. The available geometry information can be better used to obtain faster and more efficient coding with only limited loss in quality. Prediction schemes can also be applied due to the high correlation between views and between texture and depth in a drive to minimize the computation times and achieve nearer to real-time coding. Increase in SKIP modes through these techniques is also possible, further improving coding efficiency.

Scalability Issues

Dense 3D data and high quality displays will demand more and more bandwidth. Scalable representations are therefore necessary to allow for timely transmission of such data. These can be used in a number of applications, such as fast browsing, and allows adaptations to different end user devices and transmission channels. This demands the merging of multi-view video coding and scalable video coding paradigms or application of techniques such as using Wavelet-lifting-based solutions.

Holoscopic Video Coding

3D Holoscopic video is a prospective candidate format for future generation 3D TV systems. Extension of HEVC can be applied to this type of data together with efficient scaleable video coding. Given the huge amount of data, a coding platform for robust transmission of the content needs to be developed.

Error Concealment

All practical channels will disturb the transmission of data resulting in errors. Some of the errors are corrected by the error correcting codes. However, some data can still remain corrupted and needs concealment. Stereo and multi-view video concealment based on full motion field reconstruction using disparity compensated motion vector combined with traditional methods can be applied to the 3D video. Furthermore, loss in depth maps also needs to be concealed. Therefore, interpolation and contour reconstruction and geometric transforms are fundamental for the quality of the synthesized views.

Conclusion

This white paper gave an introduction to the field of 3D media coding and highlights the research direction that can bring 3D multimedia experience closer to the user. The success of the technology will allow for the development of a large number of applications and services. Europe can play an important role in this development guaranteeing jobs and economical growth. Investment in this area is clearly the way forward for 3D telemedicine applications, educational services and tutoring, entertainment, therapy, gaming, etc.

References

[1] A. Vetro, A. Tourapis, K. Müller, and T. Chen "3D-TV content storage and transmission," *IEEE Transactions on Broadcasting, Special Issue on 3D-TV Horizon: Contents, Systems and Visual Perception*, vol. 57, no. 2, pp. 384-394, Jun. 2011.

[2] Y. S. Ho, and K. J. Oh, "Overview of multimedia video coding," in *Proc. 14th Int. Workshop on Systems, Signals, & Image Processing & EURASIP Conference Focused on Speech & Image Processing, Multimedia Communications & Services,* pp. 5-12, Maribor, Slovenia, Jun. 2007.

[3] T. Chen, Y. Kashiwagi, C.S. Lim, T. Nishi, "Coding performance of Stereo High Profile for movie sequences*,"* in *ITU-T & ISO/IEC JTC1/SC29/WG11 Doc. JVT-AE022*, London, UK, 2009.

[4] ITU Video Coding Experts Group and MPEG ISO/IEC JTC1/SC29/WG11, "High Efficiency Video Coding" ITU-T Rec. H.265 and ISO/IEC 23008-2 (HEVC), 2013.

[5] C. Theobalt, G. Ziegler, M. Magnor, and H. P. Seidel, "Model-based free-viewpoint video acquisition, rendering and encoding," in *Proc. of Picture Coding Symposium*, 2004.

[6] G. Ziegler, H. P. A. Lensch, M. Magnor, and H. P. Seidel, "Multi-video compression

in texture space using 4D SPIHT," in *Proc. of IEEE Int. Workshop on Multimedia Signal Processing*, 2004.

[7] O. Boiman, S. Wurmlin, E. Lamboray, and M. Gross, "3D video fragments: Dynamic point samples for real-time freeviewpoint video," *Computers & Graphics,* vol. 28, no. 1, pp. 3-14, February 2004.

[8] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient prediction structures for multi-view video coding," *IEEE Trans. on Circuits and Systems for Video Technology,* vol. 17, no. 11, pp. 1461-1473, November 2007.

[9] A. Vetro, and D. Tian, "Analysis of 3D and multiview extensions of the emerging HEVC standard," in *Proc. of SPIE Conf. on Applications of Digital Image Processing XXXV*, Paper 8499-33, San Diego, USA, Aug. 2012.

[10] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, pp. 688-703, July 2003.

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

[12] S. Zhu, and K. –K. Ma, "A new diamond search algorithm for fast block-matching motion estimation," *IEEE Trans. on Image Processing*, vol. 9, no. 2, pp. 387-392, February 2000.

[13] J. Lu, H. Cai, J. -G. Lou, and J. Li, "An epipolar geometry-based fast disparity estimation for multiview image and video coding," *IEEE Trans. on Circuit and Systems for video Technology*, vol. 17, no. 6, pp. 735-750, June 2007.

[14] Y. T. Kim, J. Y. Kim, and K. H. Sohn, "Fast disparity and motion estimation for multiview video coding," *IEEE Trans. on Consumer Electronics*, vol. 53, no. 2, pp. 712-719, May 2007.

[15] A. Vetro, S. Yea, and A. Smolic, "Towards a 3D video format for auto-stereoscopic displays," in *Proc. of the SPIE: Appl. Digital Image Process. XXXI*, vol. 7073, September 2008.

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

[17] P. Merkle, C. Bartnik, and K. Muller, "Depth coding based on inter-component prediction of block partitition," Proc. of *Picture Coding Symposium,* pp. 149–152, 2012.

[18] G. Tech, H. Schwarz, K. Muller, and W. Thomas, "3D video coding using the synthesized view distortion change," Proc. of *Picture Coding Symposium*, pp.25–28, 2012.

Contributors:

Carl James Debono, *University of Malta*, *Malta*
Pedro A. Amado Assunção, *Instituto de Telecomunicações, Portugal*
Marco Cagnazo, *Institut TELECOM – TELECOM ParisTech, France*
Marek Domański, *Poznan University of Technology, Poland*
Sergio Faria, *Delegação de Leiria do Instituto de Telecomunicações, Portugal*
Paolo Nunes, *University Institute of Lisbon, Portugal*
Jürgen Seiler, *University of Erlangen-Nuremberg, Germany*
Mårten Sjöström, *Mid Sweden University, Sweden*