Efficient multiview depth representation based on image segmentation

Clifford De Raffaele, Kenneth P. Camilleri,
Department of Systems and Control Engineering
University of Malta
Malta
cderaffaele@ieee.org, kenneth.camilleri@um.edu.mt

Carl J. Debono and Reuben A. Farrugia
Department of Communications and Computer Engineering
University of Malta
Malta
c.debono@ieee.org, reuben.farrugia@um.edu.mt

Abstract—**The persistent improvements witnessed in multimedia production have considerably augmented users demand for immersive 3D systems. Expedient implementation of this technology however, entails the need for significant reduction in the amount of information required for representation. Depth image-based rendering algorithms have considerably reduced the number of images necessary for 3D scene reconstruction, nevertheless the compression of depth maps still poses several challenges due to the peculiar nature of the data. To this end, this paper proposes a novel depth representation methodology that exploits the intrinsic correlation present between colour intensity and depth images of a natural scene. A segmentation-based approach is implemented which decreases the amount of information necessary for transmission by a factor of 24 with respect to conventional JPEG algorithms whilst maintaining a quasi identical reconstruction quality of the 3D views.**

Keywords-3D scene processing; depth-map representation; multiview images; segmentation-based coding.

I. INTRODUCTION

Over the past years, the field of multimedia presentation has experienced relentless progression in a consistent strive to present an increasingly better quality of experience [1]. Driven by the technological progress registered in consumer electronics, as well as the substantial reduction in costs of capturing hardware equipment, customer expectations have intensified for the demand of more immersive experiences [2], [3] attained with the introduction of three-dimensional (3D) images. Realizing such a multimedia system entails the simultaneous capturing of a unique scene from multiple cameras distributed around the site.

Feasibility constraints however, dictate that its successful implementation necessitates more than advancements in acquisition hardware and progress in auto stereoscopic and holographic rendering displays. The sheer amount of data attained from capturing cameras still poses a significant challenge in the image processing and transmission domains [4]. Moreover, implementation constraints dictate that practical systems can only be achieved if the amount of instrumentation employed for scene sampling is pruned, since this, would also constitute a linear decrease in the raw video data that would require processing [5].

Depth Image Based Rendering (DIBR) techniques have demonstrated ample potential to meet this constraint. This methodology operates by generating a 3D scene from a limited number of strategically placed images around the site and their respective depth information. This depth data can be obtained either by means of specialized depth sensors employing Time-of-Flight measurements, or by utilizing stereoscopic correspondence algorithms to derive the disparity and consequently depth values for each object in the captured image frames [6]. The scene is subsequently rendered by employing Intermediate View Reconstruction (IVR) approaches that perform a 3D warp on the colour intensity image pixels by means of their respective depth values as illustrated in Figure 1. Thus, this presents the viewer with the ability to witness the scene from arbitrary perspectives within a specific range limit [7].

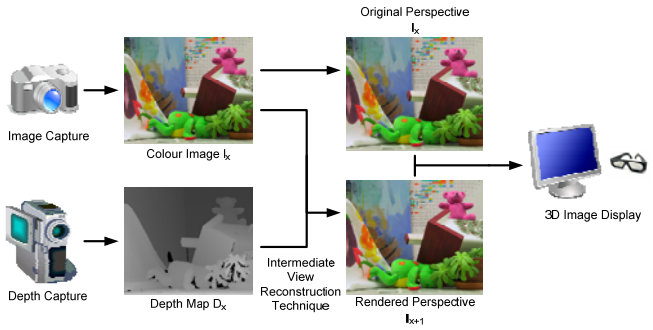

Figure 1. Block diagram overview of the depth image based rendering technique employed for 3D scene reconstruction

The pioneering approach to efficiently represent this distinctive multimedia data was devised by the European ATTEST project and involved the combination of both the image and depth data in multiview plus depth (MVD) representation format [8]. The former also indicated that the depth image can withstand further compression due to the peculiar distribution of pixel values inside the depth map. Thus, it was proposed that the standard H.264/AVC codec can be used to represent the depth data at a compression rate of 10% - 20% with respect to the colour intensity image [9]. Applying conventional image compression however, presents several artifacts on reconstructed depth images, since these techniques are particularly tuned to remove psychovisual redundancies in colour and texture whilst maintaining visual fidelity to the user [10]. Conversely, depth maps present a different situation since

these images are not directly perceived by users, but rather, employed for generating new image perspectives by IVR methodologies. The latter are especially sensitive to the preservation of data at depth discontinues for high quality rendering, whilst inexact values in uniform depth regions can be rationally endured with negligible influence [11].

These unique issues together with the distinctive properties present in depth maps have invoked substantial interest for the devising of a methodology to maintain accurate depth representation whilst exploiting inherent redundancies [12]. The image and video processing research community has thus proposed numerous techniques which range from temporal MPEG standards [13] making use of the correspondence between depth and motion vectors [14] to employ shape adaptive discrete wavelet transform (SA-DWT) for depth coding [15]. The domain of geometric modeling has also been investigated for the meticulous scenario of multiview, with depth maps being represented as voxel-based octrees [11] or by means of mesh-based coding algorithms [16]. The spatial similarity present in depth images also have been addressed by utilizing staple approaches such as JPEG-2000 and region-of-interest (ROI) coding algorithms [17]. Interesting results have also been achieved by the employment of region-based techniques [17], however these fell short of feasible implementation due to the extensive bit-rate required to accurately represent the region contours by chain-coding methods [18].

To this end, this paper proposes a novel depth coding methodology which exploits the inherent redundancies present between image and depth data for real-world scenarios. A segmentation-based approach is considered to take advantage of the spatial affinity present in depth maps and colour images alike, hence drastically reducing the amount of data necessary to represent depth maps whilst still maintaining the critical quality necessary for accurate 3D view reconstruction.

This paper is organized as follows; Section II describes the technique employed for representing depth maps together with details on the manner in which data is reconstructed at the decoder. Section III explains the implementation of the proposed algorithms and presents the obtained results. Finally, a conclusion is drawn in Section IV.

II. Depth Map Representation Technique

The approach presented in this paper takes into consideration that depth maps can be described as being composed of smoothly varying regions enclosed within sharp contours arising from object boundaries in a scene [19]. These peculiar characteristics allow the representation of depth maps to be done effectively by means of a segmentation process, since the latter would yield a number of arbitrary shaped closed regions composed of pixels with quasi homogenous values. This feature implies that coding of the depth values inside each region can be represented by singular 8-bit values portraying the respective depth of the pixels inside the area.

During decoding of the depth image however, the decoder does not have any information detailing the shape of the regions that need to be generated. Representing the exact contours of each boundary is however prohibitively expensive in terms of data, and thus would reduce the viability of the segmentation technique [18]. To this end, this paper overcomes such a restriction by proposing a novel technique which exploits the degree of correlation that exists between the colour intensity images and depth maps which are both depicting the same scene from identical viewpoints. The colour image is provided to the decoder using conventional coding algorithms. This data, present on both ends of the system, is used simultaneously to derive boundary contours by means of a segmentation process. Consequently, the representation necessary to reconstruct the depth map at the decoder is only that of a sequence of depth values obtained for the resulting regions provided as side information from the encoder.

It is here postulated that since any object captured in the depth map would present a depth boundary, the corresponding region can be rendered by at least one segment on the colour intensity image. This assumption is generally satisfied in practice, since the contours present in the depth map represent the occluding edge between a foreground object and a different background article. If the two objects have different appearance properties, such as colour and intensity values, these will be exploited by the segmentation algorithm and a boundary will be detected through image segmentation [20].

It may be noted that, if these two objects have very similar appearance characteristics, the foreground and background image regions may be merged into a single under-segmented region across the depth boundary. However, given that the two objects have distinct depth values, they are at different locations in the scene, hence they will typically be subject to dissimilar illumination from natural and artificial light sources, resulting in at least a difference in illumination dependent components. This can be clearly seen in Figure 2.

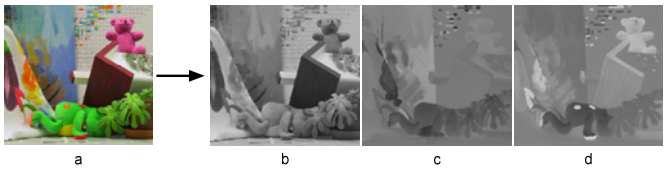

Figure 2. RGB colour image (a) is decomposed into an illumination dependent colour space before segmentation with (b) luma component, (c) blue-difference chroma component (d) red-difference chroma component

In order to ensure that objects in the scene are independently segmented, the algorithm appropriate for this requirement must be able to produce uniquely and distinctly labeled regions. Thus, approaches such as K-means clustering as well as boundary-based segmentation techniques are not suitable. The required criteria were satisfied by utilizing the Normalised Graph-Cut based segmentation [21]. This algorithm models the image I as a weighted undirected graph $G=(V,E,W)$ with each colour pixel described as a graph node V with edges E to adjacent pixels. A measure on the likelihood of pixel i and pixel j belonging to the same image region is calculated by the weight value $W(i,j)$ using [22]:

$$W(i,j) = \sqrt{W_I(i,j) \times W_C(i,j)} + \alpha WC_I(i,j) \qquad (1)$$

where the pixel affinities in intensity and intervening contours, denoted respectively by subscripts *I* and *C* are computed as:

$$W_I(i,j) = e^{-\left[\frac{\|X_i - X_j\|^2}{\sigma_X}\right]-\left[\frac{\|I_i - I_j\|^2}{\sigma_I}\right]} \quad (2)$$

$$W_C(i,j) = e^{-\frac{max_{x\in line(i,j)}\|Edge(x)\|^2}{\sigma_C}} \quad (3)$$

where X and I denote the pixel location and intensity respectively, *line(i,j)* is a straight line joining both pixels and Edge(x) is the edge strength at location x. The normalized graph-cuts subsequently partition the image into regions by performing minimum-energy cuts along the edges of the image graph [21].

III. Implementation and Results

The 3D data representation technique proposed was implemented in a manner so as to ensure backward compatibility with conventional 2D frameworks. The system thus considers the colour image to be compressed at the encoder and subsequently decoded back after the transmission or storage procedure by utilizing standard image compression codecs such as JPEG and JPEG2000 representations. To verify the proposed algorithm, the baseline *teddy* images employed were captured at a resolution of 450×375 pixels and required 157kB and 79.8kB when compressed respectively with JPEG and JPEG2000. The proposed technique hence assumes that the receiver has decoded the colour intensity data and as delineated in Section II, does not have any information relevant to the depth map prior to segmentation.

To further assist implementation feasibility as well as take advantage from the peculiar nature of the data format being decoded, the segmentation algorithm operates using the YC_bC_r color model, which represent the luminance and chroma components of the image respectively as portrayed in Figure 2. The normalized graph-cut algorithm is executed simultaneously on both the encoder and decoder terminals with parallel implementation performed for each component channel in the image. To abridge processing time and computational complexity, the algorithm is executed to perform only 50 maximum energy cuts on each image component. This data is successively combined in a conjunctive manner such that overlapping regions are fragmented into individual segments that have common support as illustrated in Figure 3.

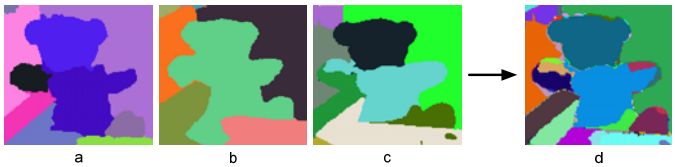

Figure 3. (a-c) Segmentations performed separately on all 3 components of the image colour-space are then utilized to derive the (d) final segmentation region map by combining the diverse partitions.

As evidenced in Figure 3d, the combined over-segmentation derived from the distinct regions in the image components ensures that the partitioning is fine enough as to describe any strong transitions in depth of the scene. Thus, depth representation is achieved by providing the decoder with a list of depth values, representing the median depth of each region. This list is provided from the encoder as side information to the compressed colour intensity image. The data is further encoded using standard Huffman coding algorithms to reduce the final size of this side information. The proposed methodology thus stipulates that the decoder, subsequent to performing an identical segmentation process, assigns to each derived region the respective depth value from the ordered side information. The reconstructed depth map, illustrated in Figure 4c, is successively employed by 3D enabled displays to generate a spatially shifted new view, which together with the original image yield a stereo image pair as portrayed in the Figure 1 procedure.

Owing to the fact that decoded depth maps are not directly viewed by users during 3D vision, comparison metrics of the depth maps fail to provide a successful quantitative representation of the quality of depth images. Thus, as described in literature [11], [13], an objective comparison of depth maps was obtained by considering the PSNR values derived from the virtual reconstructed viewpoints. These images where generated by a standard IVR algorithm [23], which as required in 3D systems, employs a decoded colour image together with its respective depth map.

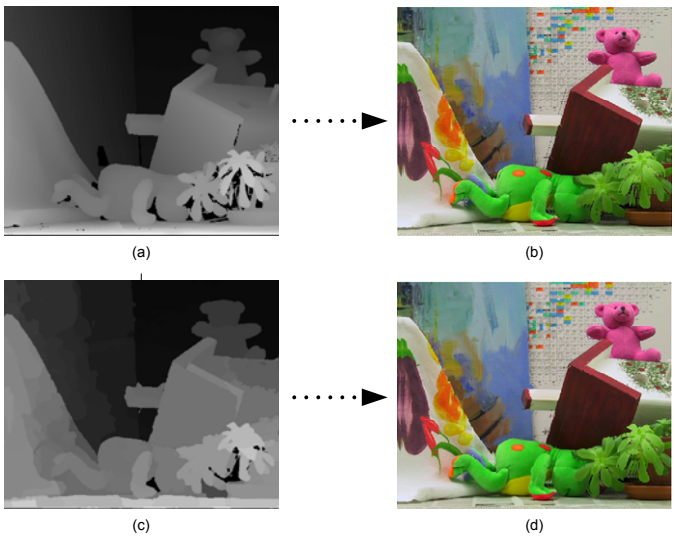

Figure 4. (a) reference uncompressed depth map, (b) reference virtual-view rendering, (c) proposed depth map representation technique, (d) virtual-view rendered utilizing the proposed depth image.

In such an *N*-camera multi-view system, a number of images I_k, where $k=1,2,...,N$, are available for a given scene such that each image I_k, represents the view from the k^{th} camera. As a baseline reference for comparison on our system, the decoded view of image I_k is rendered for the $k+1^{th}$ camera to obtain a reconstructed image I_{k+1}' using an uncompressed depth map. This process is repeated using the proposed compressed depth map to obtain I_{k+1}". Both I_{k+1}" and I_{k+1}' are then compared to the available I_{k+1} image to estimate the

rendered view quality employing the proposed depth map representation technique with respect to an uncompressed depth image as shown in Figure 4.

Observing the results in figures 4(b) and 4(d) subjectively, immediately shows that the proposed algorithm provides a rendered view quality which is comparable to that generated from an uncompressed depth map. Consistent quality results were also obtained when executing the proposed methodology to represent data in other well-known baseline images as shown in Table I. Moreover, to further expose the reduction in data rate resulting from compressing the depth image, Table I also delineates a comparison between a JPEG compressed image at conservative quality-factor of 75%.

TABLE I. PSNR AND DATA RATE COMPARISION OF DEPTH MAP REPRESENTATION TECHNIQUES

Image	Depth Representation				Rendering Quality		
	Number of Regions	*JPEG Q=75 (kB)*	*Proposed Method (kB)*	*Coding Gain (%)*	*I_{k+1}' PSNR (dB)*	*I_{k+1}" PSNR (dB)*	*PSNR Difference (dB)*
Teddy	631	12.05	0.46	2648	32.47	32.39	-0.08
Cones	636	13.40	0.50	2654	30.46	29.54	-0.92
Sawtooth	540	6.74	0.35	1948	34.67	33.53	-1.14
Average	**602**	**10.73**	**0.44**	**2417**	**32.53**	**31.82**	**-0.71**

The values expressed in Table I show that the rendered images derived from depth maps represented by the proposed technique are objectively comparable to those generated via uncompressed depth images. This is supported by the PSNR values which differ only by 0.7 dB. Moreover, this was obtained under the extreme interpolation performed by the IVR algorithm in mapping the virtual views onto a neighboring camera which would amplify discrepancies between depth images. Furthermore, the depth map compression values demonstrate that the novel methodology results in coding gain factor of 24 compared to JPEG, requiring on average only 445 bytes to represent a dense depth map of resolution 450×375.

IV. CONCLUSION

This paper has presented a novel, backward compatible, depth map representation technique to be employed for the generation of virtual viewpoints in 3D scene reconstruction. The proposed methodology employs a graph-based segmentation approach to further exploit the correspondence present between colour intensity images and depth maps captured from a unique location in natural scenes. Implementation results show that a data reduction by a factor of 24 is achieved with respect to JPEG compression algorithms whilst presenting only small losses in PSNR values for rendered images when compared to the uncompressed depth map scenario.

REFERENCES

[1] C. Fehn, P. Kauff, M. Op de Beeck, F. Ernst, W.I. Jsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, and I. Sexton, "An evolutionary and optimised approach on 3D-TV", in Proc. of Int. Broadcast Conf., pp. 357–365, Amsterdam, The Netherlands, 2002.

[2] S.U. Kum, and K. Mayer-Patel, "Intra-stream encoding for multiple depth streams", in Proc. of the 16th Int. Workshop on Network and Operating Systems Support for Digital Audio and Video, pp. 62–67, Newport, USA, May 2006.

[3] A. Kubotka, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multiview imaging and 3DTV", in IEEE Signal Process Mag., vol. 24, no. 6, pp. 10–21, Nov. 2007.

[4] A. Vetro, S. Yea, M. Zwicker, W. Matusik and H. Pfister, "Overview of Multiview Video Coding and Anti-Aliasing for 3D Displays", IEEE Int. Conf. on Image Processing, vol. 1, pp. 1-17 - I-20, Sep. 2007

[5] S.U. Yoon, E.K. Lee, S.Y. Kim, and Y.S.Ho, "A framework for multi-view video coding using layered depth images", in Proc. of the Pacific Rim Conf. on Multimedia, pp. 431–442, Jeju Island, Korea, 2005.

[6] D. Scharstein, and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms", in Int. Journal of Computer Vision, vol. 47, no. 1, pp. 7–42, Jun. 2002.

[7] S.C. Chan, H.Y. Shum, and K.T. Ng, "Image-based rendering and synthesis", in IEEE Signal Processing Magazine, vol. 24, no. 6, pp. 22–33, 2007.

[8] P. Kauff, A. Smolic, P. Eisert, C. Fehn, K. Müller, and R. Schäfer, "Data Format and Coding for Free Viewpoint Video", in Proc. of Int. Broadcast Conf., Amsterdam, The Netherlands, Sep. 2005.

[9] C. Fehn, K. Hopf, and Q. Quante, "Key technologies for an advanced 3D-TV system", in Proc. SPIE 3-DTV Video Display, pp. 66-80, 2004

[10] R. Krishnamurthy, B.B. Chai, H. Tao and S. Sethuraman, "Compression and transmission of depth maps for image-based rendering", in IEEE Int. Conf. on Image Processing, Oct. 2001.

[11] B.B. Chai, S. Sethuraman, and H.S. Sawhney, "A depth map representation for real-time transmission and view-based rendering of a dynamic 3D scene", in Proc. of 1st Int. Symposium on 3D Data Proc. Visualization and Transmission, pp. 107–114, Jun 2002.

[12] M. Flierl, A. Mavlankar, and B. Girod, "Motion and disparity compensated coding for multi-view video", in IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 11, pp. 1474 – 1484, Nov 2007.

[13] D. Tzovaras, N. Grammalidis and M.G. Strintzis, "Disparity field and depth map coding for multiview image sequence compression", in Proc. of the Int. Conf. on Image Processing, pp. 887-890, Lausanne, Sep 1996.

[14] X. Cao, Y. Liu and Q. Dai, "A flexible client-driven 3DTV system for real-time acquisition, transmission, and display of dynamic scenes", in EURASIP Journal on Advances in Signal Processing, vol. 2009.

[15] M. Maitre and M.N. Do, "Depth and depthcolour coding using shape-adaptive wavelets", in Journal of Visual Communication and Image Respresentation, v. 21, no. 5-6, 2010.

[16] S.Y. Kim and Y.S. Ho, "Mesh-based depth coding for 3D video using hierarchical decomposition of depth maps", in IEEE Int. Conf. on Image Processing, pp. 117-120, Oct. 2007.

[17] L.S. Karisson and M. Sjostrom, "Region-of-interest 3D video coding based on depth images", in Proc. of 3DTV Conf: The true vision-capture, transmission and display of 3D video, pp.141-144, May 2008.

[18] D.V.S.X. De Silva, W.A.C. Fernando and S.L.P. Yasakethu, "Object based coding of the depth maps for 3D video coding", in IEEE Trans. on Consumer Electronics, vol. 55, no. 3, pp. 1699-1706, Aug 2009.

[19] C. Cigla, X. Zabulis and A.A. Alatan, "Region-based dense depth extraction from multi-view video", in Proc of Int. Conf. on Image Processing, pp. 213-216, Texas, Sept 2007.

[20] B.Goldlucke and M.A. Magnor, "Joint 3D-reconstruction and background separation in multiple views using graph cuts", in IEEE Computer Vision and Pattern Recognition, pp. 683-688, 2003.

[21] J.Shi and J. Malik, "Normalised cuts and image segmentation", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, Aug. 2000.

[22] T. Cour, F. Benezit and J. Shi, "Spectral segmentation with multiscale graph decomposition", in IEEE Coputer Vision and Pattern Recognition, pp. 1124-1132, California, 2005.

[23] D. De Silva, W. Fernando, and H. Kodikaraarachchi, "A new mode selection technique for coding depth maps of 3D video", in IEEE Int. Conf. on Acoustics, Speech and Signal Proc., pp. 686-689. Mar. 2010