Fusion of colour and depth partitions for depth map coding

M. Maceira, J.R. Morros, J. Ruiz-Hidalgo
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
{marc.maceira, ramon.morros, j.ruiz}@upc.edu

***Abstract*—3D video coding includes the use of multiple color views and depth maps associated to each view. An adequate coding of depth maps should be adapted to the characteristics of depth maps: smooth regions and sharp edges. In this paper a segmentation-based technique is proposed for improving the depth map compression while preserving the main discontinuities that exploits the color-depth similarity of 3D video. An initial coarse depth map segmentation is used to locate the main discontinuities in depth. The resulting partition is improved by fusing a color partition. We assume that the color image is first encoded and available when the associated depth map is encoded, therefore the color partition can be segmented in the decoder without introducing any extra cost. A new segmentation criterion inspired by super-pixels techniques is proposed to obtain the color partition. Initial experimental results show similar compression efficiency to hevc with a big potential for further improvements.**

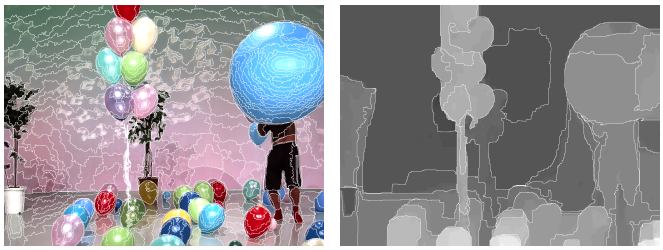
Keywords*—*Depth map coding; 3DTV; Shape-adaptive DCT; depth/texture compression;

I. Introduction

The development of three-dimensional television (3D-TV) in recent years has created the need to represent 3D visual data in a efficient fashion. The texture-plus-depth format has been widely accepted for this task. In this format, color information and depth maps (distance per-pixel samples between the camera and the 3D scene) from several viewpoints closely situated are transmitted. The use of depth information allows using depth-image-based rendering techniques at the decoder to synthesize virtual views in intermediate positions between the encoded viewpoints.

Depth maps are used to render new images and not to be viewed directly by the user. Thus, the aim when coding depth maps is to maximize the perceived visual quality of the rendered virtual color views instead of the visual characteristics of decoded depth maps themselves. Conventional image or video compression techniques have been designed for high visual quality, and are not well adapted to depth coding. The encoding of depth maps has to assess the characteristics of depth maps and reduce the transmission cost when a large number of captured views are employed [1]. An analysis of depth maps reveals that they are characterized by the presence of large homogeneous areas separated by sharp edges. Errors close to an sharp edge leads to severe rendering artifacts, while errors on areas without important transition may have negligible influence on the final quality.

There are several approaches for depth map coding. Some of them approximate the smooth changes in depth maps with piecewise-linear function in a quatree decomposition [2], [3]. Some approaches explicitly encode the position of discontinuities [4], [5], [6] to reduce the texture cost associated with depth discontinuities. Other proposals use the similarity between color information and depth maps [7], [8] to avoid encoding the location of depth map edges.



(a) (b)

Fig. 1: 1a Example of a color-based partition with 500 regions for image *Balloons*. 1b Example of a depth partition for *Balloons*.

The High Efficiency Video Coding (*hevc*) [9] standard is the most recent joint video project of the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG) standardization organizations. Enables improved compression performance relative to existing standards-in the range of 50% bit-rate reduction for equal perceptual video quality. An *hevc* extension for encoding multiple views and associated depth map [10] achieves overall savings of 40% compared with the *hevc* coding of the signals separately.

In our previous work [7], a segmentation technique was used to create a partition of the decoded color image, already available at the decoder. This partition was used to approximate the location of the depth edges, assuming that where a depth transitions occur there is probably a color contour as well. This allowed to use region-based texture coding techniques without the burden to encode the full depth partition. In this paper we present an evolution of the previous system with two main contributions:

We propose a new segmentation criterion inspired by super-pixels methods [11], [12]. Regions are combined according to the distance between its centroids. This way, compact-shaped regions, more suitable for coding purposes, are promoted.

The second contribution is related to the assumption of co-occurrence of depth and color contours in [7]. While in many cases depth and color contours are located at similar locations, there are small but noticeable differences that result in regions that contain depth discontinuities. In this case, the lack of homogeneity results in a poor performance of the region-based texture coding techniques. To solve that, we propose using a combination (fusion) of the color partition and a coarse depth partition which contains the location of the main edges, providing information about the boundaries when the color and depth discontinuities are inconsistent. In the fig. 1 a partition obtained with the proposed method for the color image (1a) and the depth partition (1b) is shown.

The paper is organized as follows. The new segmentation techniques used in this work are presented in section II. The proposed depth coding algorithm is described in Section III. In Section IV experimental results are given. Finally, concluding remarks are made in Section V.

II. Segmentation based coding techniques

A. Centroid Segmentation Criterion

The segmentation technique proposed in this work is based on the region-merging approach described in [13]. Starting from an initial partition with an arbitrary number of regions, the algorithm proceeds iteratively merging two neighbouring regions according a similarity measure or merging criterion as depicted in fig. 2. The steps in each merging step are the following:

- computing a similarity measure (merging criterion) for all pair of neighbour regions
- selecting the most similar pair of regions and merging them into a new region
- updating the neighbourhood and the similarity measures. The algorithm iterates until the desired number of regions is obtained.

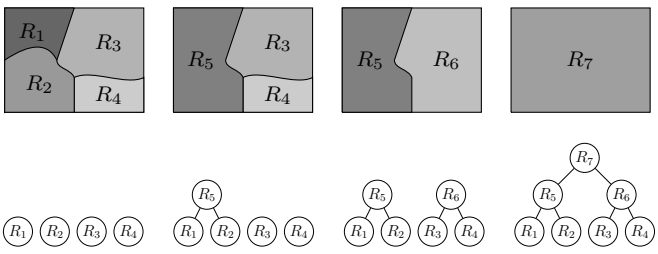


Fig. 2: From left to right, the two most-similar neighbouring regions are merged at each step. The hierarchical representation is depicted as a tree, where the region formed by merging two segments is represented as the parent of the respective nodes

In [7] the merging criterion used was the normalized weighted Euclidean distance between models, with a contour term (NWMC) described in [14]. Defining a region model constant for all the pixels within the region, the model M_R is build by averaging the values of all pixels $p \in R$, in the YCbCr color space:

$$M_R = \frac{1}{N_R} \sum_{p \in R} I(p) \tag{1}$$

where N_R is the number of pixels of region R.

The NWMC criterion consists of two terms: The first one, based on color similarity, is the weighted Euclidean distance between models (WEDM) which compares the models of the original regions with the model of the region obtained after the merging:

$$\begin{aligned} O_{WEDM}(R_1, R_2) = N_{R_1} ||M_{R_1} - M_{R_1 \cup R_2}||_2 + \\ N_{R_2} ||M_{R_2} - M_{R_1 \cup R_2}||_2 \end{aligned} \tag{2}$$

The second term is related to the contour complexity of the merged regions. The measure computes the increase in perimeter ($\Delta P(R_1, R_2)$ of the new region with respect to the largest of the two merged regions:

$$O_{Cont}(R_1, R_2) = max(0, \Delta P(R_1, R_2)) \tag{3}$$

The contour term promotes the creation of smooth contours between regions. Since most objects are regular and compact (that is, tend to have simple contours), the analysis of shape complexity can provide additional information for the mergings.

Color and contour similarity measures are linearly combined to form the NWMC criterion:

$$\begin{aligned} O_{NWMC}(R_1, R_2) = \alpha\, O_{WEDM}(R_1, R_2) + \\ (1 - \alpha)\, O_{Cont}(R_1, R_2) \end{aligned} \tag{4}$$

The NWMC criterion creates smooth regions but tends to create elongated regions which are not well suited for coding since longer regions will need more coefficients in a block-based approach. Our proposal is to add new term based on the distance between region centroids in order to obtain compact regions:

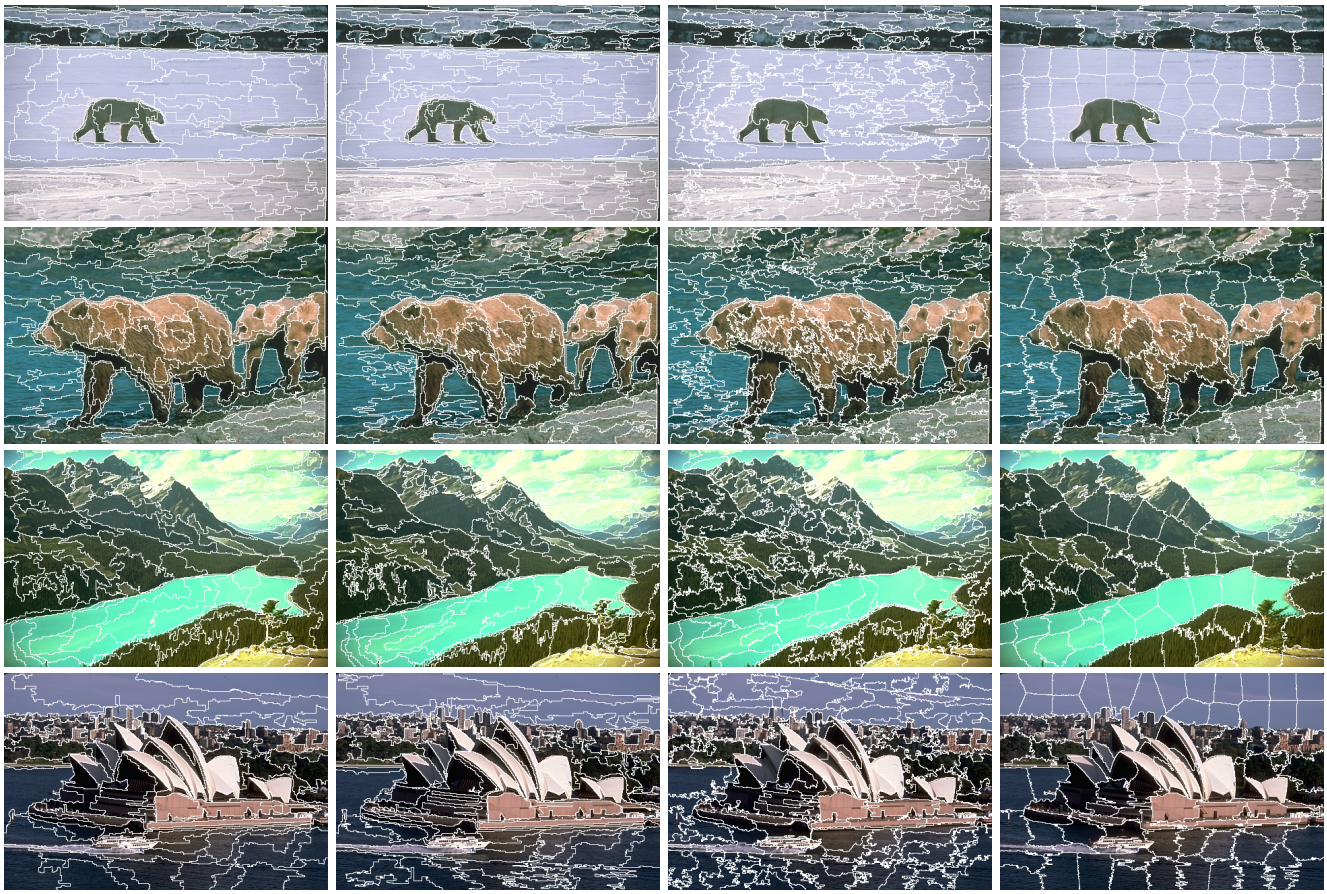
$$O_{Cent}(R_1, R_2) = \frac{1}{N_{R_1}} \sum_{p \in R_1} Coord(p) - \frac{1}{N_{R_2}} \sum_{p \in R_2} Coord(p) \tag{5}$$

After extensive testing, we found that the weight factors in (5) do not affect severely the coding performance. For simplicity, the three terms are combined with a simple addition, creating the superpixels criterion (6):

$$\begin{aligned} O_{spx}(R_1, R_2) = O_{WEDM}(R_1, R_2) + \\ O_{Cont}(R_1, R_2) + O_{Cent}(R_1, R_2) \end{aligned} \tag{6}$$

B. Depth Segmentation

The segmentation approach described in the previous section can be applied to the color image to approximate the depth partition. In [7] we assumed that a sufficiently fine partition derived from the color image should also contain the important depth boundaries. As this assumption does not always hold, a new approach is presented in this paper, combining color and depth information. It is based on merging two partitions: one constructed from the decoded color partition (available both



(a) Proposed (b) NWMC (c) seeds (d) slics

Fig. 3: Visual comparison between the various segmentation approaches used in this work. The regions generated with the proposed method present smoother contours than the NWMC method due to the centroid factor. This smoothness is desirable since most real objects have simple contours. Slics segmentation obtains even simpler contours achieving segmentations with less false boundaries at the cost of losing also some meaningful contours

at the encoder and the decoder) and one constructed from the depth map (available only at the encoder). The color partition is obtained by using the method presented in the previous section.

The depth partition is used to encode the main discontinuities of the depth map. In this case, the segmentation criterion used in the region-merging approach involves only texture information (from the depth map). We used the the squared error (SE) criterion [14] which do not take into account the size of the regions to be merged, which is preferable in the scarcely textured depth maps:

$$O_{SE}(R_1, R_2) = \sum_{p \in R_1 \cup R_2} (I(p) - M_{R_1 \cup R_2})^2 \qquad (7)$$

The partition contours are coded with a modified Freeman Chain technique [15], [16] which allows lossless coding of image partitions. Regions are represented by their boundaries and the coding process consists on tracking and encoding the boundaries. Using the fact that two consecutive boundary points in a discrete grid are neighbours, the boundaries are encoded by the movements from one contour point to another, starting from a given initial point (or various), until the complete contour is tracked.

III. Proposed Depth Map Coding Scheme

The information from color image and depth map is combined by fusing the two partitions. This fusion step creates a new partition, with more regions, which contains all the contours in both partitions. Typically, the color partition presents a more regular segmentation and also contains a good number of depth edges, while the depth partition contains the main depth edges. After the fusion process, a finer partition is obtained where the compact color partition is complemented with the depth boundaries not already present in the color image.

The scheme of the encoder process is shown in fig. 4. Color and depth map partitions are obtained independently and fused in a single image partition. As the depth partition is only available at the encoder, its contours are encoded and sent to the decoder. The fusion with the depth map segmentation solve the main inconsistencies between color and depth map, resulting in a partition with smooth variations in each region. The shape-adaptive DCT (SA-DCT) technique [17] is used to encode the texture within each region, the absence of transitions inside regions allows coding the texture information with few coefficients.

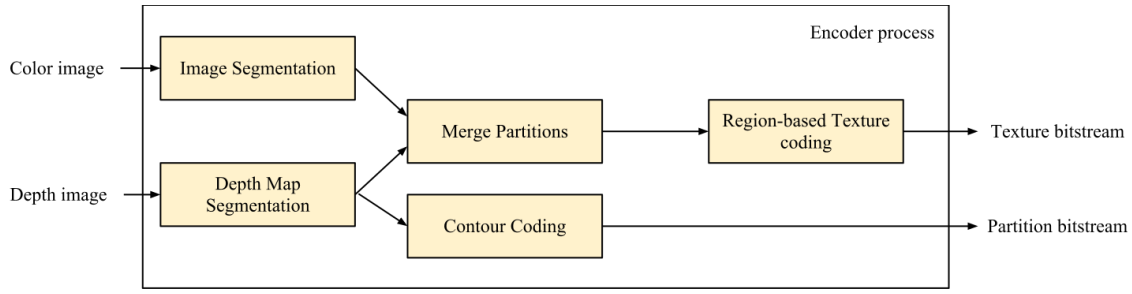


Fig. 4: Scheme for the encoding process. The two partitions obtained with the color image and the depth map are fused to a single partition which is used in the encoding of the depth map

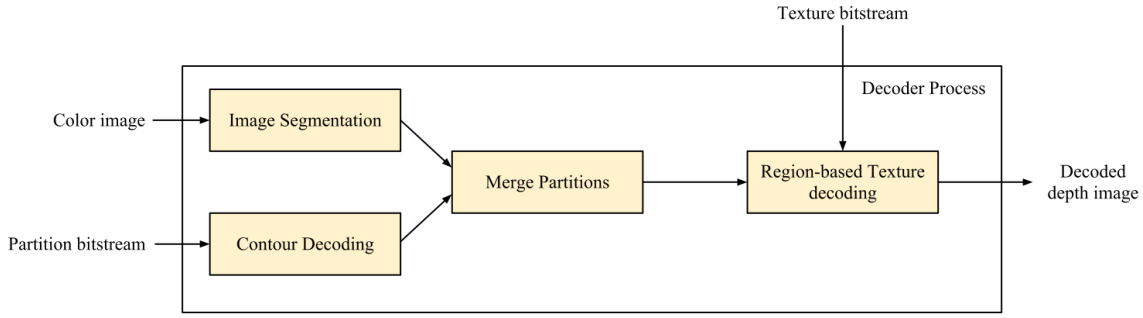


Fig. 5: Scheme for the decoding process. The color partition is obtained from the color image while the depth partition is received for the decoder

The decoder is able to replicate the color partition as it has been created from the encoder since solely the color image is used. Using the contour information received, the depth map partition is generated. The fusion segmentation is built using the same process as described for the encoder. The complete design of the decoder is presented in fig. 4.

IV. Experimental results

A. Image Segmentation results

The evaluation for this work has been performed in two separated environments. Firstly, a comparison with other superpixels segmentations is performed to validate the use of the new centroid distance. The benchmark for this work consist in the BSDS500 [18] which contain 500 images with human-marked-boundaries as ground-truth. A comparison with our previous segmentation technique and with two state of the art methods *-seeds* [11] and *slics* [12]- is provided. In fig. 3 a visual comparison between the diverse segmentation approaches used in this work is provided.

The fig. 6 shows the numerical results in terms of precision, recall and f-measure for boundaries between segmentation and ground-truth. In the proposed method, the use of the centroid decreases the recall with respect to the NWMC criteria but the precision is improved obtaining a performance similar to the segmentation obtained with *slics*. Globally the usage of the centroid criteria achieves a better trade-off between superpixel compactness and boundary adherence than NWMC. The loss of precision in the boundaries is compensated in the depth coding technique proposed in this work by the depth map segmentation.

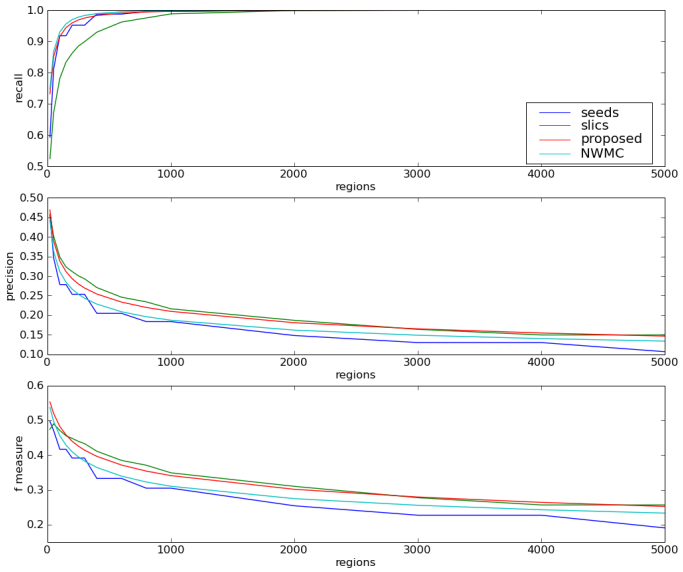
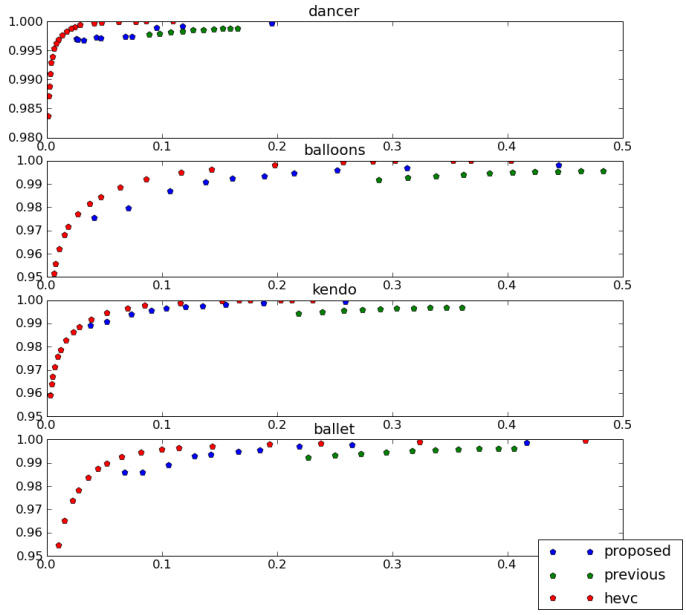


Fig. 6: Precision, recall and f-measure depending on the number of regions

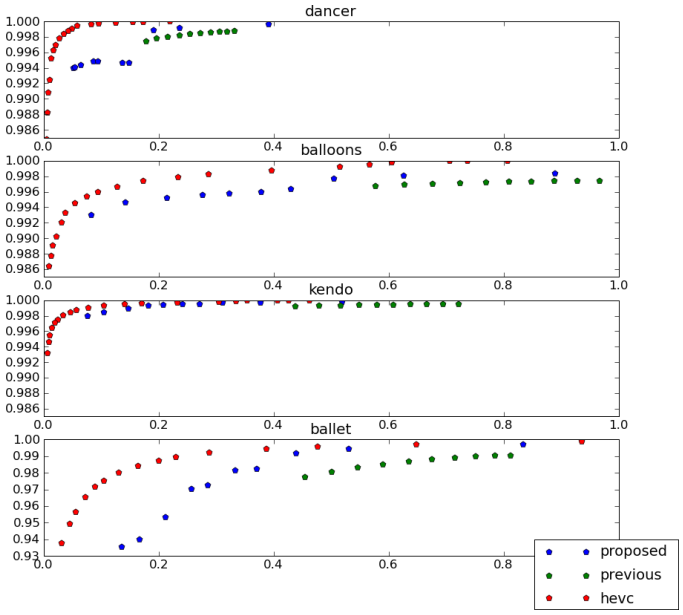
The f-measure results are comparable to *slics* with the advantage of obtaining a hierarchy of regions. This hierarchy allows a rate-distortion optimization that could be used in the coding process.

B. Depth Map coding results

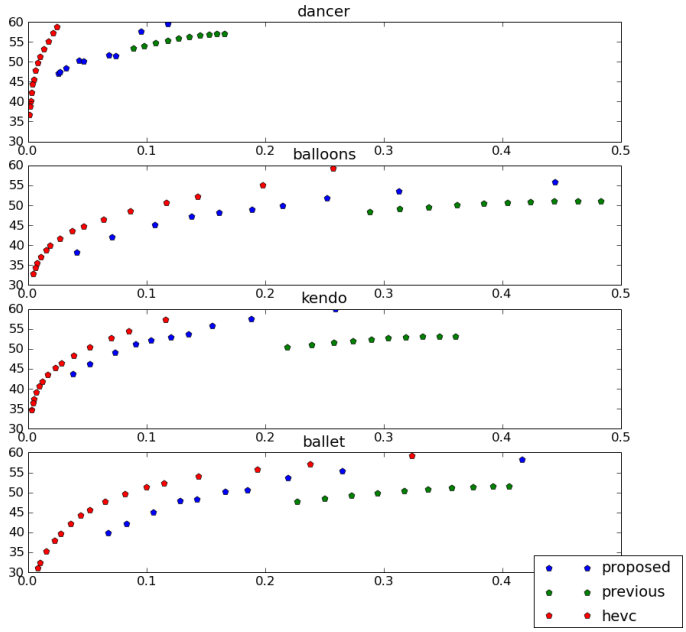
The proposed coding method is evaluated using 10 images of each of the multiview sequences sets *ballet*, *Undo Dancer*,



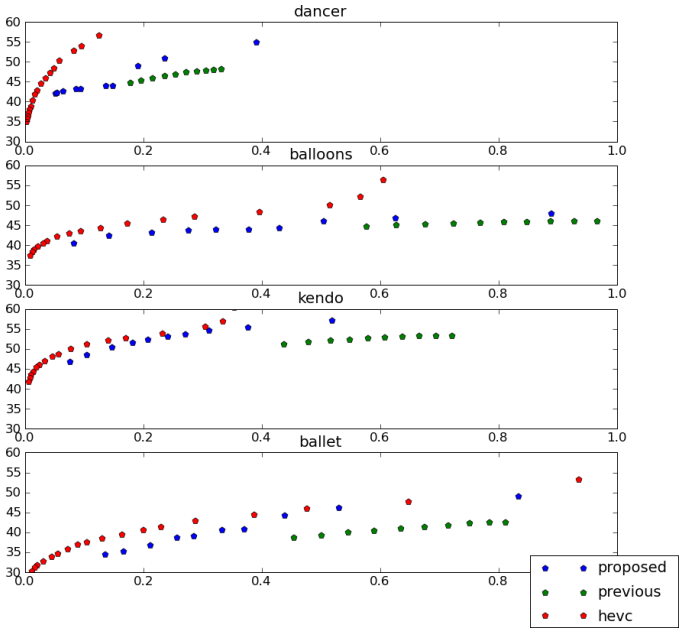
(a) SSIM over depth map



(b) SSIM over virtual views



(c) PSNR over depth map



(d) PSNR over virtual views

Fig. 7: Structural Similarity (SSIM) and Peak signal-to-noise ratio (PSNR) over depth maps and in virtual views of our method compared with hevc and our previous approach for different number of bits per pixel

	dancer	balloons	kendo	ballet
Depth Map				
Previous	-12.39	-10.93	-12.76	-8.79
Proposed	-15.81	-6.94	-6.70	-5.00
Virtual View				
Previous	-8.43	-6.48	-7.91	-4.97
Proposed	-12.66	-4.46	-3.42	-2.96

TABLE I: Bjontegaard measures with respect to hevc

kendo and *balloons*. As our algorithm is image based, only intra-mode comparison with the main intra-mode profile *hevc* is provided. The method is tested for a fixed number of regions in the segmentations of 50 regions in the depth segmentation and 100 regions in the color segmentation. The number the regions in the depth segmentation has been found to be sufficient in finding the main edges on the image. The number of regions in the color partition can be increased in order to obtain higher distortion figures at the cost of augmenting the total rate as the number the regions coded increases.

As the aim of the depth map is synthesize new images, the quality of the method is evaluated directly over the depth map but especially in the intermediate view. The intermediate view is synthesized with the view synthesis reference software VSRS [19] using the original color images and the compressed depth maps. Error measures are computed between the synthesized view with the uncompressed depth maps and with depth maps compressed with the method proposed, with *hevc* and with our previous proposal [7].

The results are presented in both peak signal-to-noise ratio (psnr) and structural similarity (ssim) depending on the number of bits per pixel used to build the depth map image: encode the contours of the depth segmentation and encode the texture coefficients. The cost of encode the contours given the depth segmentation is fixed whereas the texture of the fusion segmentation is encoded using different quantization steps. The results can be seen in fig. 7 and the bjontengaard [20] results for the psnr are summarized in table I. The bjontengaard figures are presented with respect to *hevc*.

In both ssim and psnr the proposals presented in this work outperforms our previous approach. The bjontengaard improvement is in the margin of 3 dB except for the dancer sequence where for low rates the quality measures does not improve diminishing the bjontengaard values for that sequence. In our previous work the rate-distortion optimization prevented that behaviour.

The results obtained are below *hevc* but the segmentation techniques proposed in this work reduce the margin with the state of the art. The advantages of a segmentation-based technique can be seen in the better performance obtained with our technique in virtual views. Tests using only depth map segmentation show a further reduction in texture cost encouraging further research in segmentation-based depth map coding.

V. Conclusion

A novel depth map coding algorithm is proposed in this paper. Using the redundancy between the color view and depth map, a segmentation over the color image is used to approximate the boundaries in the associated depth map. Additionally a segmentation is done in the depth map directly to solve inconsistencies between color image and depth map. The fusion of both segmentations separate the smooth areas in depth maps reducing the texture coding cost. Experimental results using shape adaptive dct on the generated regions shows an improvement over our previous work especially measured in virtual views where the advantages of region-based texture coding are exhibited. Further investigations will examine the further improvement of using this new methodology in combination with rate-distortion optimization over a hierarchy of regions.

Acknowledgeable

This work has been partially supported by the Ministry of Education of Spain (FPI grant BES-2011-045678) and the Spanish Ministerio de Ciencia e Innovación, under project TEC2010-18094

References

[1] P. Merkle et al., "Multi-view video plus depth representation and coding," in *ICIP*, Oct 2007.

[2] Y. Morvan, P. de With, and D. Farin, "Platelet-based coding of depth maps for the transmission of multiview images," in *Proceedings of SPIE: Stereoscopic Displays and Applications*, vol. 6055, 2006.

[3] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Muller, P. de With, and T. Wiegand, "The effect of depth compression on multiview rendering quality," in *3DTV Conference*, May 2008, pp. 245–248.

[4] M. Maitre and M. Do, "Shape-adaptivewavelet encoding of depth maps," in *PCS*, May 2009, pp. 1–4.

[5] F. Jager, "Contour-based segmentation and coding for depth map compression," in *VCIP 2011*, Nov. 2011, pp. 1–4.

[6] I. Daribo, G. Cheung, and D. Florencio, "Arithmetic edge coding for arbitrarily shaped sub-block motion prediction in depth video compression," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, 30 2012-oct. 3 2012, pp. 1541 –1544.

[7] M. Maceira, J. Ruiz-Hidalgo, and J. Morros, "Depth map coding based on a optimal hierarchical region representation," in *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2012*, oct. 2012, pp. 1 –4.

[8] H. Deng, L. Yu, J. Qiu, and J. Zhang, "A joint texture/depth edge-directed up-sampling algorithm for depth map coding," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, july 2012, pp. 646 –650.

[9] G. Sullivan, J. Ohm, W.-J. Han, T. Wiegand, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1649 –1668, dec. 2012.

[10] H. Schwarz, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, K. Muller, H. Rhee, G. Tech, M. Winken, D. Marpe, and T. Wiegand, "Extension of high efficiency video coding (hevc) for multiview video and depth data," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, 30 2012-oct. 3 2012, pp. 205 –208.

[11] M. Bergh, X. Boix, G. Roig, B. Capitani, and L. Gool, "Seeds: Superpixels extracted via energy-driven sampling," vol. 7578, pp. 13–26, 2012.

[12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274 –2282, nov. 2012.

[13] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval," *IEEE Trans. on IP*, vol. 9, no. 4, pp. 561–576, Apr. 2000.

[14] V. Vilaplana, F. Marqués, and P. Salembier, "Binary partition trees for object detection," *IEEE Trans. on IP*, vol. 17, no. 11, pp. 2201–2216, Nov. 2008.

[15] H. Freeman, "On the coding of arbitrary geometric configurations," *IRE Trans. Electronic, Comp.*, p. EC(10):260268, June 1961.

[16] F. Marqués, J. Sauleda, and T. Gasull, "Shape and location coding for contour images," in *Picture Coding Symposium*, Lausanne, Switzerland, March 1993, pp. 18.6.1–18.6.2.

[17] T. Sikora and B. Makai, "Shape-adaptive dct for generic coding of video," *IEEE Trans. on CSVT*, vol. 5, no. 1, pp. 59–62, Feb. 1995.

[18] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int'l Conf. Computer Vision*, vol. 2, July 2001, pp. 416–423.

[19] "View synthesis reference software (vsrs) 3.5," ISO/IEC JTC1/SC29/WG11, Tech. Rep., 2010.

[20] G. Bjontegaard, "Calculation of average psnr differences between rd-curves," *ITU-T SG.16 Q.6, Document VCEG-M33*, April 2001.