DEPTH MAP CODING BASED ON A OPTIMAL HIERARCHICAL REGION REPRESENTATION

M. Maceira , J. Ruiz-Hidalgo , J.R. Morros

Department of Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC)
{*marc.maceira, j.ruiz, ramon.morros*}*@upc.edu*

ABSTRACT

Multiview color information used jointly with depth maps is a widespread technique for 3D video. Using this depth information, 3D functionalities such as free view point video can be provided by means of depth-image-based rendering techniques. In this paper, a new technique to encode depth maps is proposed. Based on the usually smooth structure and the sharp edges of depth map, our proposal segments the depth map into homogeneous regions of arbitrary shape and encodes the contents of these regions using different texture coding strategies. An optimal lagrangian approach is applied to the hierarchical region representation provided by our segmentation technique. This approach automatically selects the best encoding strategy for each region and the optimal partition to encode the depth map. To avoid the high coding costs of coding the resulting partition, a prediction is made using the associated decoded color image.

Index Terms — Depth map coding, 3DTV, Shape-adaptive DCT, depth/texture compresion, rate-distortion optimization

1. INTRODUCTION

Three-dimensional television (3D-TV) is becoming one of the main developments in visual home entertainment experience. With recent development of 3D displays and interactive applications, 3D-TV has become a point of interest by academics and industries. To enable 3D-TV applications, complementary information of the 3D scene are required, such as video captured from different viewpoints and one corresponding depth map for each viewpoint. The depth map consist of depth samples associated at each pixel from the video frame, encoded as a gray level image. As the amount of data required in 3D-TV is typically very large, depth map information must be efficiently compressed. Because depth maps are used to render new images (virtual views) and not viewed directly by the end user, the goal when coding depth maps is to maximize the perceived visual quality of the rendered virtual color views instead of the visual characteristics of the depth maps themselves.

The straightforward solution to compress depth map images with conventional image or video compression algorithms may not be appropriate since traditional image compression methods have been designed to provide maximum perceived visual quality. The rendering error depends on the quality of the depth map coding and the quality of the coding of the original color view used as a reference, but also on the structure of the depth map. Depth maps are characterized by having homogeneous areas separated by sharp edges. Errors in the depth map close to an sharp edge can result in severe rendering artifacts, while errors on a smooth area may have negligible influence on the final quality. Therefore, encoding the edges with fidelity is indispensable.

The encoding of depth maps of a large number of captured views [1] leads to a high transmission cost. Several solutions for Multi-View Video coding have been proposed, with and without temporal prediction. Among these without temporal prediction are [2, 3] based on a quadtree decomposition that divides the image into blocks of variable size, each block being approximated by one platelet (piecewise-linear function) or [4, 5] where schemes with edge extraction and explicit signaling the location of discontinuities were proposed. Other proposals modify existing block algorithms by adding a new coding mode able to handle partitions inside each block. In [6], an alternative mode to MVC is proposed where color information is used to construct two different regions inside the block. In [7] and [8], modifications to the intra prediction mode of H.264/AVC are proposed by creating regions in the blocks and edge-depending prediction for each region.

Our proposal has some similarities with these methods as it also inspired in segmentation-based coding systems but, in our case, the entire depth map image is segmented and each of the regions is separately encoded. We propose to use an optimal lagrangian approach to jointly create the final segmentation and select the most efficient (in a rate-distortion sense) texture coding strategy. The paper is organized as follows. The proposed algorithm is described in Section 2. In Section 3 experimental results are given. Finally, concluding remarks are made in Section 4.

2. PROPOSED ALGORITHM

Our proposal is based on segmenting the depth map into homogeneous regions of arbitrary shape and then coding the contents of these regions using texture coding techniques. Having uniform regions will allow coding the texture inside the region with only a few

coefficients. However, as the receiver does not know the shape of the resulting regions, the partition must be encoded and transmitted. To avoid the high cost associated to coding the resulting partitions (region shape), instead of directly segmenting the depth map and sending the partition to the receiver, an approximate depth partition is constructed using the decoded color image that is supposed to be available at the receiver.

This work is an extension of our previous approach [9] that already used combined color and depth information to construct the final partition. We propose to use an optimal lagrangian approach applied to the hierarchical region representation provided by our segmentation technique. The lagrangian approach presents two benefits. Firstly, the encoding parameters are selected optimally for each region and secondly, the optimal final number of regions needed to encode the depth map can be automatically obtained. Furthermore, the previous texture coding technique based on polynomial basis functions has been replaced in this work by a shape adaptive discrete cosine transform [10] which allows to obtain a better rate-distortion trade-off.

Next sub-section explains the segmentation technique used in this work. The segmentation technique allows us to create a hierarchical region representation that will be later used in the lagrangian optimization. This process is done combining the information of the depth map to encode and the associated decoded color image thus allowing the decoder to re-create the same segmentation in the decoding step.

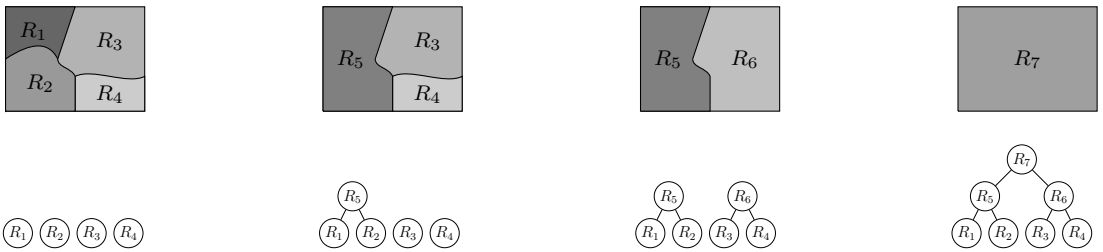


Figure 1: BPT creation: From left to right, the two most-similar neighbouring regions are merged at each step. The BPT representation is depicted as a tree, where the region formed by merging two segments is represented as the parent of the respective nodes

2.1. Color-depth hierarchical representation

Our segmentation technique is based on a region-merging approach. Starting from an initial partition with a high number of regions (or even starting at the pixel level), the hierarchical representation is constructed using an iterative process: at each iteration, the two more similar regions, defined by a *merging criterion* O, are merged until only one region, the root node, is reached. This iterative process allows construct a hierarchy of regions named binary partition tree (BPT) [11]. Fig. 1 shows a simple bpt creation process. By pruning the resulting hierarchical representation at a given level, we can obtain a partition with any desired number of regions.

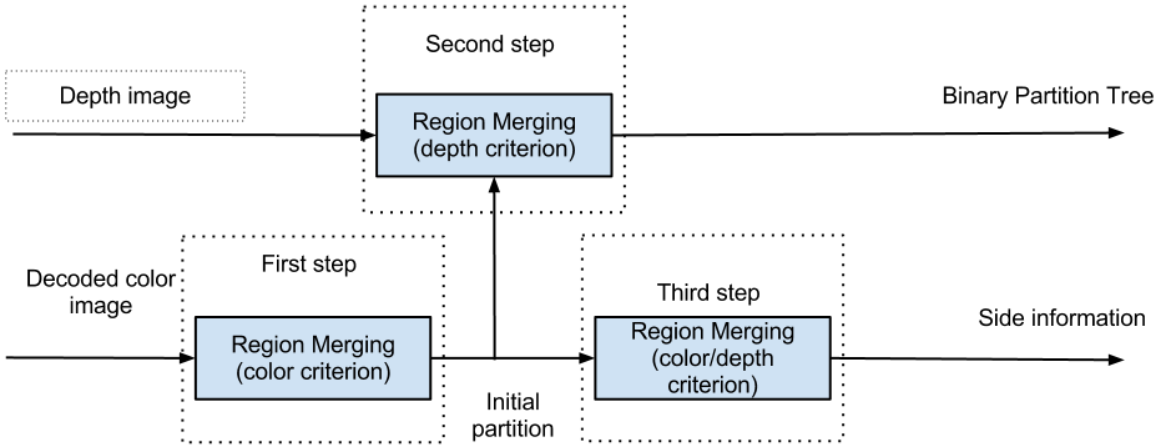
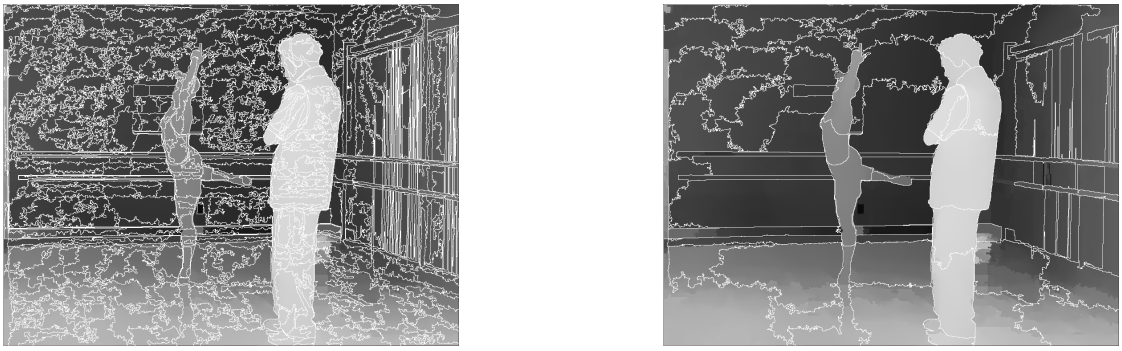


Figure 2: Block diagram for the hierarchical representation

The BPT creation process is depicted in Fig. 2 and can be described in three steps:

First step: A fine initial color partition (initial partition) is built in the same way both at the encoder and the decoder (using the decoded color view). This initial partition can be considered an approximation of the final depth partition. The assumption here is that depth transitions coincide in most situations with color transitions. Since the color fine partition can be constructed both at the encoder



(a) (b)

Figure 3: 3a Example of a color-based initial partition with 500 regions for image *Ballet*. 3b Example of a final partition selected by the proposed Lagrange optimization for image *Ballet*.

at the decoder, it has no coding cost. Experiments have shown that our approach is not sensible to the number of regions in this initial partition and simply, a high enough number to prevent under-segmentation is enough (Fig. 3a).

Second step: The BPT is obtained at the encoder side by using a merging criterion applied to the depth information (O_{depth}), starting from the initial partition constructed in the previous step. Using an initial partition constructed using only depth information (in the first step) would be optimal, but in that case, the shape of this partition would need to be sent, at a high coding cost.

Third step: As the second step uses depth information (not available at the decoder) to evaluate the merging criterion, a third step is needed to be able to recreate the same segmentation on the decoder, by creating some side information. The idea is to create another BPT using a merging criterion applied to the decoded color information (O_{color}). In this case the merging steps order will differ with respect to the BPT created using O_{depth}, but, since the starting fine partition is the same in both cases and depth transitions usually coincide with color transitions, there will be a great amount of similarity between both trees and in a lot of cases, the result provided by O_{depth} and O_{color} will coincide. In this manner, the decoder can recreate the same (depth-based) BPT of the second step using the initial partition, the O_{color} merging criterion and the side information (that indicates whether O_{depth} and O_{color} differ). This side information can be encoded using one bit per merging (we send a '1' when criteria differ and '0' when coincide). Our experiments have shown that using this approach instead of directly encode the region shapes greatly reduces the amount of bits needed to encode the partition. The amount of bits needed to replicate the hierarchical structure is limited to less than 0.01 bits per pixel in all images used in our experiments.

2.2. Lagrangian multiplier optimization in BPT

Lagrangian multiplier optimization [12] is an effective technique to evaluate a large number of possible coding choices in an optimization fashion. In this work we will use lagrangian minimization to find the best partition inside the BPT and the set of quantizers and the texture coding strategy (coding technique) for each region taking into account the hierarchical structure of the BPT.

The optimization relies on the technique discussed in [13]. The problem can be formulated as the minimization of the distortion D of the image with the restriction that the total cost R be below a given budget. It has been shown that this problem can be reformulated as the minimization of the Lagrangian: $D + \lambda R$ where λ is the so-called Lagrange parameter. Both problems have the same solution if we find λ^* such that R is equal (or very close) to the budget. Therefore, the problem consists in using the BPT tree to find a set of regions creating a partition and a set of coding techniques minimizing $D + \lambda * R$ (Fig. 4).

Figure 4: Lagrangian optimization over BPT. The bottom-up analysis follows the merging sequence from Fig. 1 activating the nodes with minimum lagrangian. The algorithm finds the best partition (regions 3, 4 and 5) and the best coding strategy.

Assuming in a first step that the optimum λ^* is known, the first step is to make a local analysis and to compute, for each node of the BPT, the Lagrangian for each coding technique. The coding technique giving the minimum Lagrangian is considered as the optimum one for this node and this Lagrangian is stored. The second step is to define the best partition. This can be done by a bottom-up analysis of the BPT. Starting from the lowest level, one checks if it is better to code the area represented by the children regions as a single region

X or as individual regions $\{x_i\}_{i=1,2}$ with $X = \cup x_i$. The best choice is selected by comparing the Lagrangian of X with the sum of the Lagrangians of x_i

$$d(X) + \lambda r(X) < \sum_i d(x_i) + \lambda r(x_i) \quad (1)$$

If the former is lower than the latter, the node corresponding to X is activated and the children nodes are deactivated. This inequality is evaluated following the merging sequence up to the root node. Note that, to use this approach, the distortion should be additive over the regions. In our experiments, the squared error has been used; however, any additive measure can be used. At the end of the procedure, the best partition is defined by picking up all the regions corresponding to the activated nodes together with their corresponding best coding technique (defined during the first step of the algorithm). The definition of the optimum parameter λ can be done with a gradient search algorithm.

Fig. 3b shows an example of the partition selected by the lagrangian optimization for the image *Ballet*. At the end, the total encoding rate for all regions is formed by the sum of each of the regions texture cost $r_t(X)$ plus the coding cost of sending the side information that will allow the decoder to build the same BPT.

2.3. Texture coding strategies: Shape-Adaptive DCT

The lagrangian multiplier optimization presented in the previous section can select, among a set of different coding techniques (a specific combination of a texture coding strategy and a quantizer), the optimum one for each of the segmented regions. In this work, a shape-adaptive discrete cosine transform (sa-dct) [10] has been used as the texture coding strategy. The sa-dct allows to encode the texture (in this case the depth associated with each pixels of a region) of any arbitrary shaped segment by means of separate the 2D-dct segment transform in two 1D-dct transforms.

Two modes of the sa-dct has been used. The first mode, sa-dct_{blocks}, divides the segmented region in blocks of $8x8$ pixels and each block is encoded using sa-dct or standard dct. The sa-dct is used in blocks that correspond to a boundary of the region while the standard dct is used in blocks where all pixels belong to the region. The second mode, sa-dct_{single}, is intended to handle homogeneous regions in an efficient manner, hence encodes all the segmented region in a single sa-dct giving a rough approximation to the texture with few coefficients.

Both modes are available to the lagrangian multiplier optimization process that selects the optimal one, region by region, following the optimization described in the previous section. Once the optimization process finds the optimal partition the texture coefficients from active regions are grouped and entropy coded using an adaptive arithmetic codec.

3. EXPERIMENTAL RESULTS

The proposed method is evaluated using the depth map images of two multiview sequence sets: *Ballet* and *Book arrival*. As our algorithm is image based, only a single frame of the sequence is employed. Furthermore, comparisons are provided against still image coding methods, such as JPEG and JPEG2000.

The presented results are obtained with initial partitions for the encoding process of 2000 regions. Experimental tests show that this number do not affect the overall system performance since the lagrangian optimization procedure automatically finds the optimal number of regions for the final partition.

The quality of the depth map coding can be evaluated in two ways: a) directly over the depth map and b) over synthesized virtual views. As the purpose of depth maps is to create virtual views, quality comparison using the synthesized views is the fairest method to evaluate the depth map compression. Both evaluations are presented in this Section. First, a comparison directly over depth map quality is shown in Fig. 5. Second, a comparison over virtual views extracted using the MPEG view synthesis reference software (VSRS 3.5) is studied in Fig. 6.

The left plot of Fig. 5 shows quality comparison over depth maps with different configurations, sa-dct_{single} and sa-dct_{blocks} correspond to encode the initial partition with one coder whereas the other approaches uses the proposed lagrangian optimization. The lagrangian optimization finds the best trade-off for each region improving the use of a single coding for the whole image. This can be evaluated by the improvement seen with respect to the *lagrangian no hierarchical* curve that the final number of regions has been fixed manually. In addition, the lagrangian optimization allows encoding the image with a mixture of sa-dct_{single} for homogeneous regions and sa-dct_{blocks} in regions with higher texture variations.

The right plot of Fig. 5 shows a comparison of the proposed method with JPEG and JPEG2000. In this case, the PSNR has been measured directly over the encoded depth maps. In terms of DB-PSNR the gain is 0.86 dB to JPEG and -5.7 dB JPEG2000. The lower performance of our proposed method was expected as the encoding of the region texture is not as precise as JPEG2000. However, contours are better preserved (being a segmentation based coding system) and with higher virtual view quality.

In order to validate the assumption that our method provides virtual views with higher quality, an intermediate view is synthesized with the VSRS reference software using the original color images and the compressed depth maps. PSNR measure is computed between the synthesized view with the uncompressed depth maps and with depth maps compressed with our approach and with JPEG and JPEG2000. Fig. 6 shows the rate-distortion curves for both images. The BD-PSNR over the virtual views is 5.83 dB to JPEG and -1.05dB to JPEG2000 for *Ballet* and 3.43 dB to JPEG and 1.4 dB to JPEG2000 for *Book arrival*. The Fig. 7 shows to synthesized views with the same bitrate. The performance of our method greatly outperforms JPEG and it is slightly lower than JPEG2000.

4. CONCLUSION

In this paper a novel depth map coding algorithm is proposed. Exploiting the redundancy between the color view and depth map, a hierarchical region based model, a BPT, is used to separate the smooth areas in the depth maps allowing a reduced texture coding cost. A lagrangian algorithm selects the texture coding strategy for each region and the final partition. Moreover, experimental results using shape adaptive dct transforms are provided leading to gains up to 5.8 dB in PSNR in synthesized view quality with respect to JPEG and similar results as JPEG2000. Further investigations will explore wavelets encoding for the texture as well as exploit temporal and inter-view redundancy.

5. REFERENCES

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

[2] Y. Morvan, P.H.N. 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*, 2006, vol. 6055.

[3] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Muller, P.H.N. 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.N. 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] S. Liu, P. Lai, D. Tian, C. Gomila, and C.W. Chen, "Sparse dyadic mode for depth map compression," in *ICIP*, Sep. 2010, pp. 3421–3424.

[7] G. Shen, W. Kim, S.K. Narang, A. Ortega, J. Jaejoon Lee, and H. Wey, "Edge-adaptive transforms for efficient depth map coding," in *PCS*, Dec. 2010, pp. 566–569.

[8] B.T. Oh, H. Wey, and D. Park, "Plane segmentation based intra prediction for depth map coding," in *PCS*, may 2012, pp. 41–44.

[9] J. Ruiz-Hidalgo, J.R. Morros, P. Aflaki, F. Calderero, and F. Marqués, "Multiview depth coding based on combined color/depth segmentation," *Journal of Visual Communication and Image Representation*, vol. 23, pp. 42–52, 2011.

[10] 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.

[11] 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.

[12] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *Signal Processing Magazine*, vol. 15, no. 6, pp. 23–50, Nov. 1998.

[13] P. Salembier et al., "Segmentation-based video coding system allowing the manipulation of objects," *IEEE Trans. on CSVT*, vol. 7, no. 1, pp. 60–74, Feb. 1997.

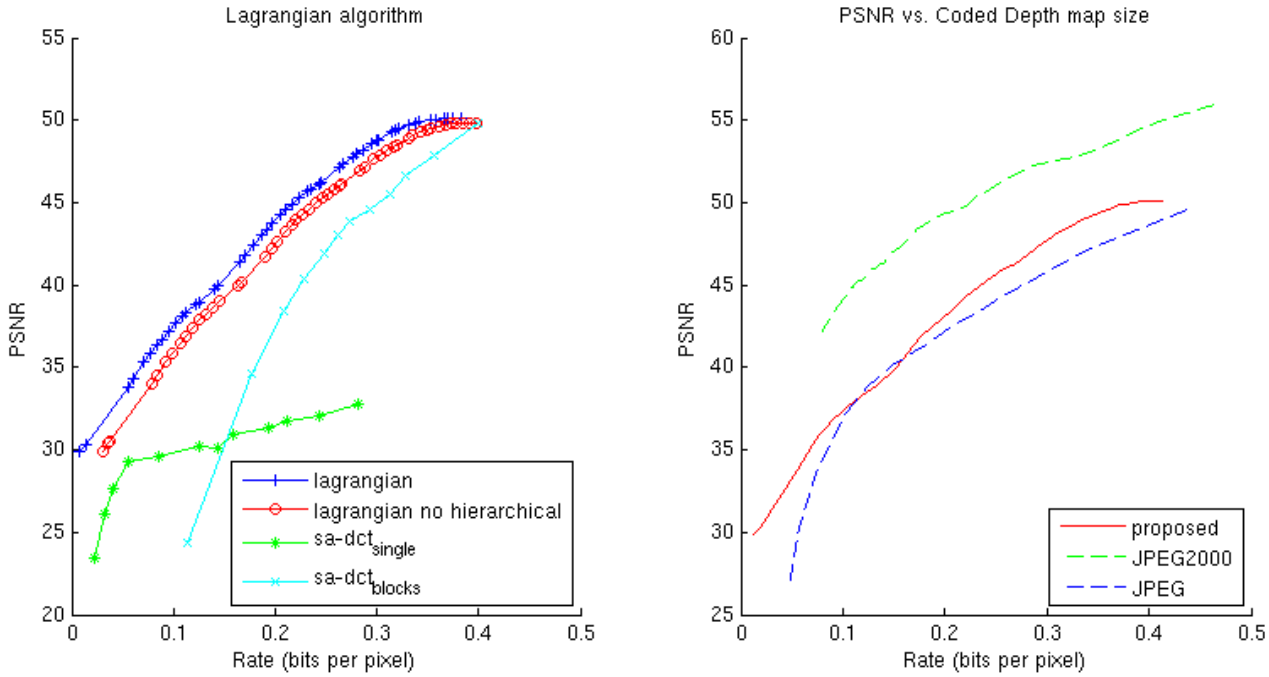


Figure 5: Rate-distortion over depth maps. On the left the proposed lagrangian optimization compared with single texture coding strategies. On the right, rate-distortion performance of our method compared with JPEG and JPEG2000.

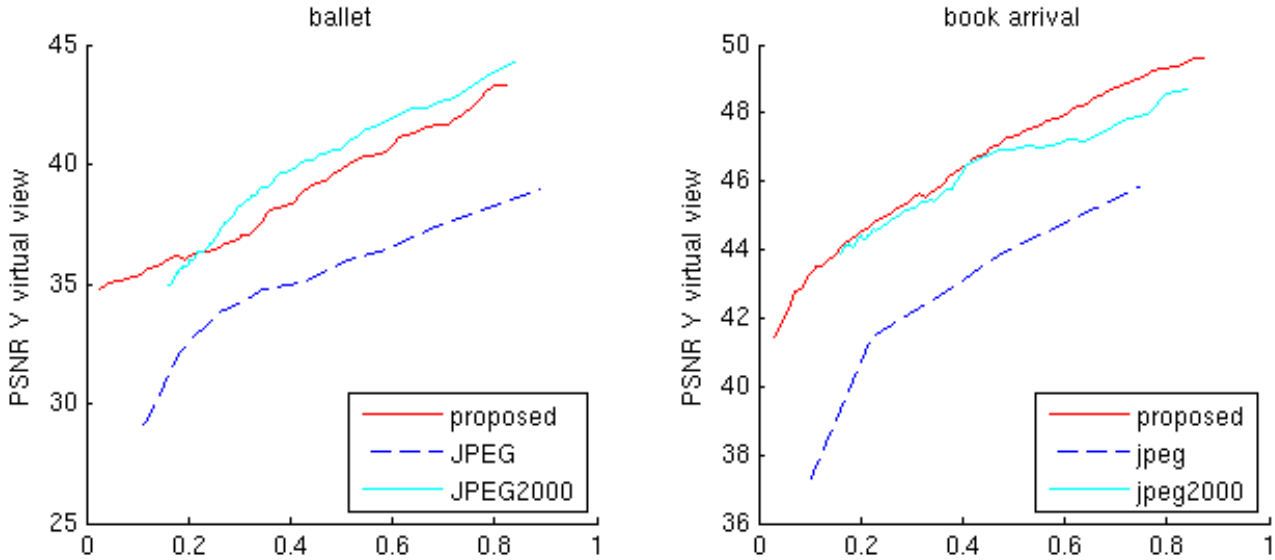


Figure 6: Rate-distortion curves over virtual view for *Ballet* and *Book arrival*. Depth images coded with JPEG, JPEG2000 and with the proposed algorithm

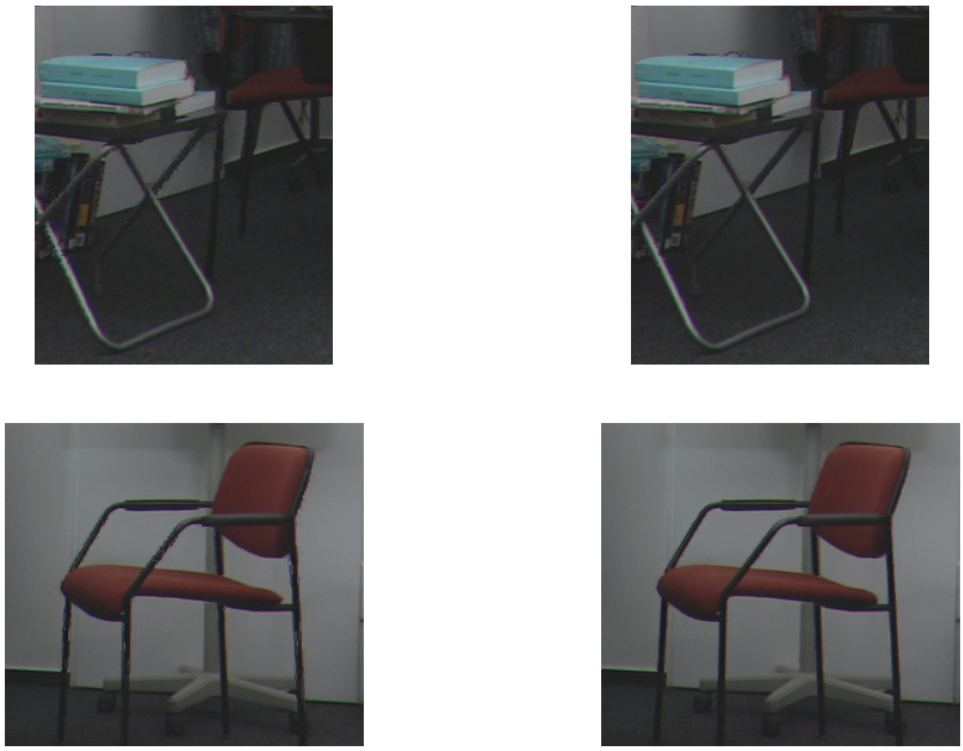


Figure 7: Visual synthesis results for image *Book arrival* using depth map coded at 0.2 bits per pixel. Left synthesis using JPEG, right using the proposed method.