A JOINT 3D IMAGE SEMANTIC SEGMENTATION and SCALABLE CODING SCHEME with ROI APPROACH

Khouloud Samrouth, Olivier Deforges, Yi Liu, Wassim Falou, Khalil Mohamad

A JOINT 3D IMAGE SEMANTIC SEGMENTATION and SCALABLE CODING SCHEME with ROI APPROACH

Khouloud Samrouth #1, Olivier Deforges #2, Yi Liu #3, Wassim Falou *4, Mohamad Khalil *5

UEB, CNRS UMR 6164, IETR Lab, INSA de Rennes
20, avenue des Buttes de Coesmes, CS 70839, 35708 Rennes, France
[1] khouloud.samrouth1@insa-rennes.fr
[2] olivier.deforges@insa-rennes.fr; [3] yi.liu@insa-rennes.fr

* LaSTRe, EDST, Lebanese University
Mitein Street, Tripoli, Lebanon
[4] wassim.falou@ul.edu.lb; [5] mohamad.khalil@ul.edu.lb

***Abstract*—Along with the digital evolution, image post-production and indexing have become one of the most advanced and desired services in the lossless 3D image domain. The 3D context provides a significant gain in terms of semantics for scene representation. However, it also induces many drawbacks including monitoring visual degradation of compressed 3D image (especially upon edges), and increased complexity for scene representation. In this paper, we propose a semantic region representation and a scalable coding scheme. First, the semantic region representation scheme is based on a low resolution version of the 3D image. It provides the possibility to segment the image according to a desirable balance between 2D and depth. Second, the scalable coding scheme consists in selecting a number of regions as a Region of Interest (RoI), based on the region representation, in order to be refined at a higher bit-rate. Experiments show that the proposed scheme provides a high coherence between texture, depth and regions and ensures an efficient solution to the problems of compression and scene representation in the 3D image domain.**

***Index Terms*—3D scalable compression, segmentation, RoI extraction and refinement**

I. Introduction

Recent technologies in multimedia systems focus on the post-production and indexing services. On the one hand, the post-production allows editing the scene by changing the real background to a virtual one [1]. Recent High Definition movies (*HD movies*) such as *Avatar*, use the post-production service to replace the real background with a virtual environment rapidly, accurately and automatically. Thus, such a feature becomes a desirable service in the film-making domain. On the other hand, the indexing, consisting of both auto-extraction and objects retrieval in the scene, is an interesting and effective solution for visual management of image databases such as digital libraries, remote sensing and medical imaging [2], [3].

978-1-4799-6139-9/14/$31.00

Such services offer a satisfying experience for users in the professional domain as they operate on lossless-coded images. Today, post-production and indexing are also widely used in the public domain to edit and reconstruct taped scenes. However, images and videos in the public domain are coded in a lossy way, and the compression distortions can affect the reliability of the scene representation.

The extension from 2D to 3D for such applications provides significant gain in terms of semantics for scene representation. However, it also brings two main drawbacks. Firstly, the subjective quality of a compressed 3D image, generally coded as 2D+Z (depth), will be very dependent on the distortion upon edges. Secondly, the introduction of a new space dimension significantly increases the complexity for scene representation.

In this paper, we do not propose a new 3D image segmentation method such as [4], [5], [6], nor a new 2D+Z coding such as in [7], but rather a joint and a complementary coding and region-level representation scheme. The main interest is to globally consider the problems of compression and representation in order to preserve the consistency between 2D (texture) image, the depth image and regions.

The first part of the paper (Sections II and III) presents this global coding scheme, based on a coder called LAR (Locally Adaptive Resolution). LAR is an initially lossy 2D coding framework established on a content-based QuadTree partition [8]. Recent extensions to scalable lossy/lossless, introduced in [9], proposed a joint texture and depth image compression solution. In [10], combined compression and region representation coding methods are proposed. These methods consists of performing first the compression of the 2D image, and then the segmentation starting from the QuadTree partition. The first part of the contribution presented in this paper is an extension to 3D context of the work of Sekkal et al. [10], with a semantic segmentation. The term semantic refers to the possibility to adjust region-level representation according to the balance between 2D and depth defined by users, as illustrated in Fig. 1: if only color is taken into account for segmentation, then

objects of different colors have to be segmented in different regions. If only depth is taken into account, then only the objects belonging to the same depth layer will be in the same region, even if they have different colors. A balance between color and depth can be used in order to obtain a suitable segmentation. whatever the semantic configuration, the region representation preserves the main contours of the objects with pixel accuracy.

The second part of the paper (Section IV) introduces an extension of the proposed coding/representation scheme to RoI (Region of Interest) coding. The underlying paradigm is to decode at a higher bit-rate the RoI, selected from segmented regions of low bit-rate compressed 3D images. The general idea is to decode a low bit-rate compressed 3D image, perform the segmentation, select a number of regions, and finally decode at a higher bit-rate the RoI (still with pixel accuracy).

This paper is organized as follows. In Section II, the LAR coder framework and the global coding scheme are briefly described. In Section III, the proposed 3D semantic segmentation and its results are presented. In Section IV, we present the extension to RoI coding. Finally, we conclude this paper in Section V.

a) b) c) d) e)

Fig. 1. a) Texture, b) Depth, c,d,e) Semantic segmentation according to the balance between 2D and depth.

II. LAR CODER FRAMEWORK

A. 2D LAR coder

Locally Adaptive Resolution (LAR) [8] is an efficient multiresolution content-based 2D image coder for both lossless and lossy image compressions (see Fig. 2).

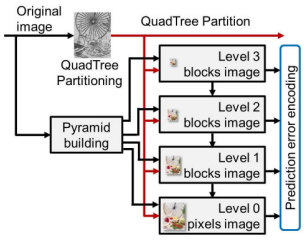
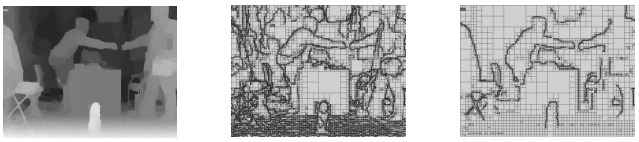


Fig. 2. LAR multiresolution coder scheme

The LAR coder relies on a local analysis of image activity, leading to a QuadTree representation with variable-size block partitioning ($1 \times 1, 2 \times 2, ..., 64 \times 64, 128 \times 128$). The activity analysis stage is based on an edge detection criterion so that the smallest blocks are located upon edges, while large blocks map homogeneous areas (Fig. 3). Then, the main feature of the LAR coder at low bit-rates is to preserve contours while smoothing homogeneous parts of the image. The coding method consists in representing the input image at its QuadTree level, and encoding associated values through successive dyadic decompositions, transforms and predictions stages. The coder involves two input parameters: Q_p the quantization parameter for the prediction error, and the Th_{Quad} the threshold for the edge criterion. Th_{Quad} is generally set to $\frac{2}{3}Q_p$, which is a near optimal value for all images. Then, $Q_p = 1$ means lossless coding.



a) Original image b) $Th_{Quad} = 5$ c)$Th_{Quad} = 20$

Fig. 3. Examples of QuadTree partition of a) depth image Bookarrival View 10 Frame 33 for threshold b) Th_{Quad} = 5, c) Th_{Quad} = 20.

B. Global 2D+Z coding scheme

A joint depth/texture scalable coding method has been proposed in [9], (cf. global scheme in Fig. 4). Firstly, the 2D image is encoded at low bit-rate considering only Z information to build the QuadTree. Secondly, the Z image is encoded on the same QuadTree, with an improved prediction stage using the Y component of the previously coded 2D image. Thirdly, 2D quality improvement can be obtained by refining the QuadTree partition considering both 2D and Z images. This joint coding method preserves the spatial coherence between texture and depth images, and reduces visual artefacts for synthesized views, especially upon edges. The coding scheme also has a reduced complexity in comparison with JPEGXR, and outperforms it in terms of objective compression efficiency [9].

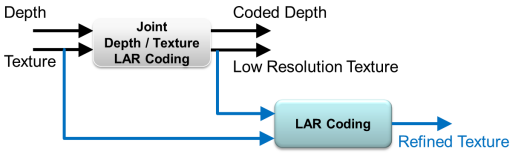


Fig. 4. Global 2D+Z coding scheme

III. 3D SEGMENTATION

A. Semantic Approach of Region Representation

In order to preserve the coherence between a region-level representation and the 2D image coding technique, authors proposed in [8] a segmentation process started from the QuadTree partition, and followed by successive region fusions. The segmentation process is based on a Region Adjacency Graph (RAG) representation, and directly acts on blocks instead of pixels to limit the complexity.

We propose here an extension of this 2D segmentation to 3D context. The main feature in a segmentation process is the homogeneous criterion. It is generally defined by a cost function initially considered between two regions R_i and R_j. Here, the cost function $Cost_I(R_i, R_j)$ in image I is function of the global means of the two regions and the local gradient of adjacent blocks of the two regions.

The segmentation process obviously requires as input images the texture, the depth and the QuadTree. However, several choices exist: QuadTree based on depth only, on texture only,

or on both? Texture image at low or high resolution? In the context of low bit-rates compression and coarse regions representation, the QuadTree based on depth only ($Quad_Z$) is preferred, along with the low bit-rate /resolution compressed texture (LR) and depth for the same partition. In the opposite context, the QuadTree based on depth and texture ($Quad_{ZT}$) will be used with high bit-rates texture (HR) and depth.

The semantic of interest, in terms of 2D and/or depth information, will be application-dependent. This semantic is introduced in the segmentation process as input parameters for the luminance (Y), chrominances (Cb, Cr), and depth (Z). Then, the segmentation cost between 2 regions is computed using a weighting mix of texture and depth information:

$$Cost(R_i, R_j) = \alpha \times Cost_Y(R_i, R_j) + \beta_1 \times Cost_{Cb}(R_i, R_j) + \beta_2 \times Cost_Cr(R_i, R_j) + \gamma \times Cost_Z(R_i, R_j).$$

with $\alpha,\ \beta_1,\ \beta_2$ and γ being the input weighting coefficients representing the semantic of interest.

The most common fusion criterion used in segmentation methods is the minimal cost between two regions. This type of solutions is generally considered as optimal, but requires large amount of computation. To speed up the process, we propose a fusion criterion based on a threshold: a cost is firstly calculated between a region and each of its adjacent regions, and then the merge operation is allowed if the minimum of these costs is below a given threshold (Th_{Seg}). This parameter controls the granularity of the final region representation: the greater the Th_{Seg}, the lower number of regions we will have as more regions will be merged.

The depth image generally contains much less information than the luminance component of the texture. The consequence is that for the same Th_{Seg}, the number of regions issued from depth only will be small compared to the segmentation from the texture only. To resolve the problem and have a comparable level of representation, we introduce an adaptive threshold solution. Considering that any increase in rate demands a higher fusion threshold, the relative compression rates (R_k) of each component is integrated. For a given $Th_{SegInput}$, the expression of the new threshold Th_{Seg} is:

$$Th_{Seg} = \frac{\alpha.R_Y + \beta_1.R_{Cb} + \beta_2.R_{Cr} + \gamma.R_Z}{\alpha + \beta_1 + \beta_2 + \gamma} Th_{SegInput} \quad (1)$$

$$\text{where } R_k = \frac{rate_k}{rate_Z};\ k = Y,\ C_b,\ C_r,\ Z;$$

B. Results and Discussion of Region Representation

At this step, comparisons with existing techniques are not feasible. Indeed, to the best of our knowledge, the proposed global representation and coding is unique in terms of combined functionalities. For instance, the comparison with advanced state of the art methods for 3D segmentation has no sense since the segmentation is performed at low resolution at block-level instead of pixel-level. From compression point of view, comparisons can be found in [9] with standard coding techniques, but in the proposed scheme requiring fine grain representation, no existing coding standard could replace the proposed coding scheme.

The proposed semantic segmentation approach has been successfully tested on a large set of MPEG 3D reference sequences (*Balloons, BookArrival, Newspaper, UndoDancer*). The codec contains different parameters : Th_{Quad} for the quality of the compressed 2D+Z image, $(\alpha, \beta_1, \beta_2, \gamma)$ for the semantic choice between 2D and Z, and $Th_{SegInput}$ for the merging criteria, and finally the choice between a Quadtree based on the depth only or on depth and texture. Therefore, we can show only a few examples among all possibilities in these experiments. For the following results, we set Th_{Quad} = 33, with different $Th_{SegInput}$ and different combinations of weighting coefficients.

Figure 5 and Figure 6 show results in a low bit-rate context and a high bit-rate context, respectively for Undodancer example. In Fig. 5, the segmentation process starts from the $Quad_Z$ only (low bit-rate context), which speeds up the segmentation. The results of segmentation are presented in false colours, the mean of depth per region and the mean of luminance per region, in each column respectively. In Fig. 5.d, the image is partitioned into 285 regions while in Fig. 5.e, where only depth is considered, the image is coarsely partitioned into 60 regions. Figures 5.f and 5.g show good compromises between depth and texture information.

Figure 6 shows a higher bit-rate context considering the $Quad_{ZT}$ with a finer region representation. In Fig. 6.d, where only HR texture is considered, the image is partitioned into a higher number of regions and the dancer is partially merged with the background. Two examples, in which the HR texture and the HR depth are considered, are shown in Fig. 6.e and Fig. 6.f. We can notice that the image is partitioned into lower number of regions, and has a more reliable representation.

This approach provides several interests: 1) it is possible to obtain the suitable region representation from very eye-compressed images by balancing between 2D and depth according to the semantic of interest of the user or application; 2) it is possible to switch from a segmentation process with detailed scene representation to a fast process with global scene representation; 3) whatever the semantic configurations, the proposed region representation method preserves the main contours of the objects with pixel accuracy, and maintains a spatial coherence between depth and texture images.

IV. REFINEMENT USING RoI APPROACH

Four stages are performed in this application. Firstly, the decoder decodes a low bit-rate compressed 2D+Z image, using a QuadTree based on depth only ($Quad_Z$). Secondly, the segmentation is performed at both coder and decoder side. Thirdly, a RoI is selected at the coder or decoder side, in a semi-automatic or automatic way (for instance, with a 3D object segmentation method). Fourthly, the refinement for the texture is inside the RoI only and is realized by the decoder, by locally refining the QuadTree partition based on depth and texture.

The RoI is a sub-set of the $Quad_Z$, and this QuadTree itself is a sub-set of the $Quad_{ZT}$. Therefore, the RoI can always match the refined QuadTree partition, and the local refinement is easy and direct: it only consists in firstly extracting a binary mask composed of blocks in the initial QuadTree partition ($Quad_Z$), and then using this mask for the validation of the local refinement partition.

The quantization factors Q_p can be independently set to any value during the first and second coding stage. Fig. 7 gives an example of results. One can notice that the RoI is refined with a noticeable visual quality improvement. Thus, such extension can be used for accurate local object enhancement or extraction.

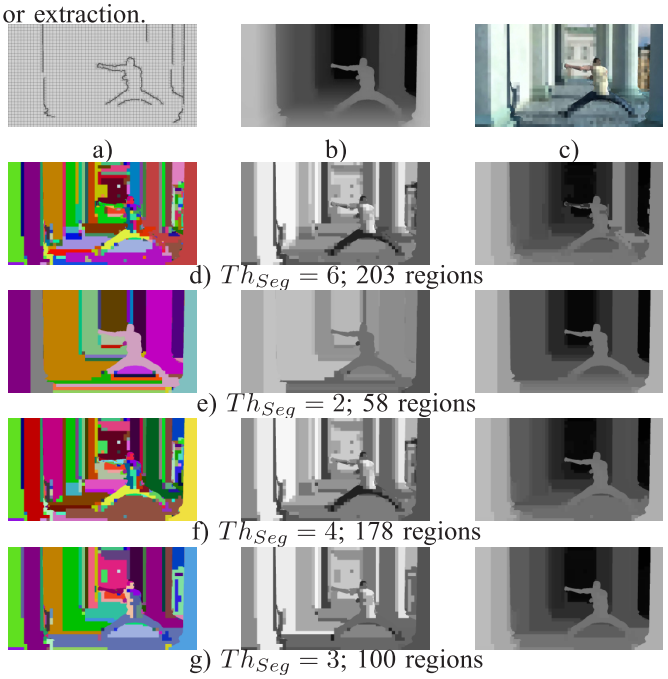


Fig. 5. Low Resolution Segmentation results of Undodancer view 1 frame 250 (1888x1024 pixels) using the a) $Quad_Z$ at 0.003 bpp, b) depth coded at 0.017 bpp (38 dB) and c) texture coded at 0.04 bpp (17 dB). Segmentation based d) on texture($\alpha = 1, \beta_1 = \beta_2 = \gamma = 0$), g) 80% on depth and 20% on texture ($\alpha = 0.2, \beta_1 = \beta_2 = 0, \gamma = 0.8$).

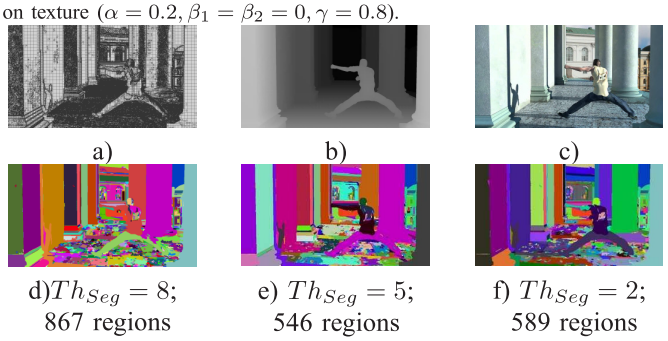


Fig. 6. High Resolution Segmentation results of Undodancer view 1 frame 250 using the a) $Quad_{ZT}$ at 0.04 bpp, b) depth coded at 0.03 bpp (44.26 dB) and c) texture coded at 0.41 bpp (31.95 dB). Segmentation based d) on texture($\alpha = 1, \beta_1 = \beta_2 = \gamma = 0$), e) 50% on texture and 50% on depth ($\alpha = 0.5, \beta_1 = \beta_2 = 0, \gamma = 0.5$), f) 80% on depth and 20% on texture ($\alpha = 0.2, \beta_1 = \beta_2 = 0, \gamma = 0.8$).

V. Conclusion

This paper presents a unique full coding system dedicated to 3D images for compression, region representation and RoI extraction /refinement. In this approach, the 3D image is firstly encoded by the scalable LAR coder. Then, a segmentation, which can be guided by the 3D semantic, is performed from low bit-rate images. The proposed region representation preserves the main contours of the objects with pixel accuracy. Finally, a local refinement is realized for a RoI deduced from the region representation. Results showed that the method can accurately capture the content of the scene, with a refinement precisely located inside the RoI. Future work will focus on an automatic extraction of RoI from the semantic region representation.

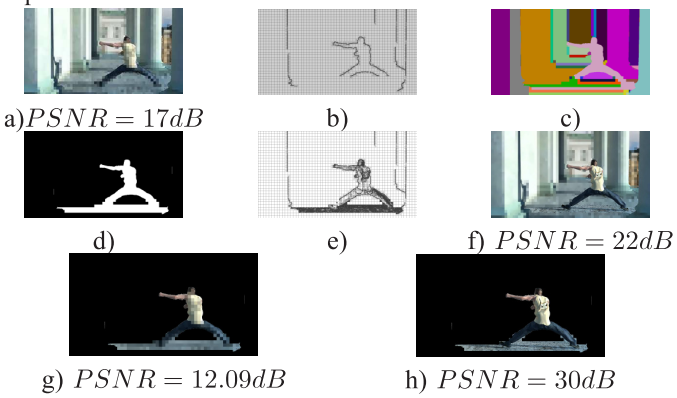


Fig. 7. Refinement results of Undodancer view 1 frame 250. a) LR Texture at 0.04 bpp ($Q_p = 50$), b) $Quad_Z$ ($Th_{Quad} = 33$) at 0.003 bpp, c) Segmentation partition d) Binary Mask of RoI, e) Refined QuadTree at 0.03 bpp f) Refined texture at 0.8 bpp, g) LR and Extracted RoI, h) Refined and Extracted RoI.

References

[1] D. Corrigan, F. Pitie, V. Morris, A. Rankin, M. Linnane, G. Kearney, M. Gorzel, M. O'Dea, C. Lee, and A. Kokaram, "A video database for the development of stereo-3d post-production algorithms," *Conference on Visual Media Production (CVMP)*, pp. 64–73, 2010.

[2] F. Zargari, A. Mosleh, and M. Ghanbari, "Compressed domain jpeg2000 image indexing method employing full packet header information," *International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 410–416, 2008.

[3] Y. Wu and J. Liu, "Image indexing in dct domain," *International Conference on Information Technology and Applications (ICITA)*, vol. 2, pp. 401–406, 2005.

[4] C. Cigla and A. Alatan, "Segmentation in multi-view video via color, depth and motion cues," *IEEE International Conference on Image Processing (ICIP)*, pp. 2724–2727, 2008.

[5] H. He, D. Mckinnon, M. Warren, and B. Upcroft, "Graphcut-based interactive segmentation using colour and depth cues," *Australasian Conference on Robotics and Automation*, 2010.

[6] X. Dai, "Automatic segmentation fusing color and depth," *International Conference on Pattern Recognition (ICPR)*, pp. 763–766, 2012.

[7] J. Fu, D. Miao, W. Yu, S. Wang, Y. Lu, and S. Li, "Kinect-like depth data compression," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1340–1352, Oct 2013.

[8] O. Deforges, M. Babel, L. Bedat, and J. Ronsin, "Color lar codec: A color image representation and compression scheme based on local resolution adjustment and self-extracting region representation," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 17, no. 8, pp. 974–987, 2007.

[9] K. Samrouth, O. Deforges, Y. Liu, F. Pasteau, M. Khalil, and W. Falou, "Efficient depth map compression exploiting correlation with texture data in multiresolution predictive image coders," *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6, 2013.

[10] R. Sekkal, C. Strauss, F. Pasteau, M. Babel, and O. Deforges, "Fast pseudo-semantic segmentation for joint region-based hierarchical and multiresolution representation," *SPIE Electronic Imaging - Visual Communications and Image Processing*, Jan. 2012.