Scene Segmentation by Color and Depth Information and its Applications

Carlo Dal Mutto Pietro Zanuttigh Guido M. Cortelazzo

Department of Information Engineering
University of Padova
Via Gradenigo 6/B, Padova, Italy
{dalmutto,pietro.zanuttigh,corte}@dei.unipd.it

Abstract

Scene segmentation is a classical problem in computer vision and image processing. Traditionally this problem has been tackled by exploiting only color information from a view of the scene. Recently, a great amount of research has been done in order to develop efficient solutions for depth estimation and tools like modern stereo vision techniques and ToF range cameras allow accurate depth information extraction. The availability of both color and depth information opens the way for new insights in scene segmentation algorithms. In this paper, a novel scene segmentation framework is introduced. The goal of this approach is to capture the scene distribution, taking into account both color and depth cues, in order to provide a more accurate segmentation. Numerous applications can benefit from an effective scene segmentation method, e.g. 3D video, free-viewpoint video and game controlling. In this paper we will also show how the proposed segmentation scheme can be used to improve depth data compression performances, by using it in order to identify and extract the edges and the main objects in the scene.

1. Introduction

Image segmentation is a very challenging task that has been the subject of a huge amount of research activity. There are many different segmentation techniques based on different tools. A first class of methods is based on graph theory and in particular on the formulation of the segmentation problem in terms of a graph-cut problem [2]. Another group of methods are based on clustering algorithms, e.g. the method of [1] exploits the mean shift clustering algorithm for image segmentation. There are also methods based on region merging, level sets, watershed transforms, spectral methods and many other different techniques. Despite the great effort spent on this issue, it remains an unsolved problem. One of the main reasons is that the information content of the image is not always sufficient to properly recognize the object in the framed scenes (e.g. consider an object over a background of the same color).

In the last years, many different solutions have been proposed for the extraction of depth information relative to a real world scene. Various systems have been proposed in order to solve this task, each one with pros and cons. There are two main group of methods, passive and active methods. Passive methods use only the information coming from two or more standard cameras to estimate the depth. Among them stereo vision systems, that exploit the localization of corresponding locations in the different images, are perhaps the most widely used. A complete review of this family of methods can be found in [6]. Stereo vision systems have been greatly improved in the last years but they can not work on untextured regions and the most effective methods are also very computational time consuming. Another possible solutions are the active methods such as structured light, laser scanners and Tof sensors [3]. By projecting some form of light on the scene such methods can obtain better results than passive stereo vision systems, but they are also usually more expensive.

Depth data can be segmented easier than color images and allows to recognize objects that have similar colors but at the same time close objects with similar depths can not be easily identified (e.g. two close people side-by-side). It is quite common to have both depth and color data relative to the same framed scene and by exploiting both the geometry (depth) information and the color clues it is possible to better recognize the objects in the scene and thus improve segmentation performances. This paper follows this rationale and introduces a novel segmentation algorithm that exploit both depth and color information. In Section 2 we describe the joint depth and color representation of the samples and we introduce the proposed segmentation algorithm. Section 3 presents the experimental results and shows how the joint segmentation algorithm allows better performance than us-

ing color or depth alone. In Section 4 we show how the proposed segmentation scheme can be used for depth data compression and finally in Section 5 we draw the conclusions.

2. Proposed segmentation algorithm

Scene segmentation is the process of analyzing, understanding and labeling the various parts that compose a scene. The proposed segmentation algorithm assumes the availability of both geometrical and color information about a scene.

This assumption leads to the fact that given a view of a generic scene $\mathcal{S}$, for each point $\mathbf{p}_i \in \mathcal{S}, i = 1, ..., N$ (i.e. each pixel in the acquired image), it is possible to get its geometrical position (defined by its 3D coordinates in the camera reference frame $[x, y, z]^T$) and its color value $[R, G, B]$.

While the color value of a point comes directly from the color image of the framed scene, the full geometrical position can be derived from the depth value and the image coordinates of the point $\mathbf{p}$, by simply inverting the camera projection matrix.

The input of our segmentation algorithm are:

- the color image C of the scene $\mathcal{S}$, acquired by a standard camera
- the depth-map D of the framed scene $\mathcal{S}$ (it could have been acquired by any 3D reconstruction method, e.g. a ToF camera or a stereo vision system)
- the projection matrix relative to the acquired color image and depth-map

2.1. Color image

In order to exploit in the best way the color information, it is important to consider a color representation that is well suited for the segmentation task. In particular we choose to adopt a uniform color space, because this kind of color space preserve the distances between the perceived colors in the representation space. In this way the distances used in the clustering process of Section 2.3 are consistent with the perceived color difference. Note also that a uniform color space also ensures that distances in each of the 3 color components are consistent, thus making easier to perform the clustering of the 3-vector representing the color information. In particular the CIELab color representation system has been used for this work. So, for each point $\mathbf{p} \in \mathcal{S}$, a 3-vector is defined in order to account for the color information in the CIELab color space:

$$\mathbf{p_i^c} = \begin{bmatrix} L(\mathbf{p}_i) \\ a(\mathbf{p}_i) \\ b(\mathbf{p}_i) \end{bmatrix} \tag{1}$$

2.2. Computing geometric information

While for the color image, the extraction of the information is a well known problem, the extraction of the geometric information for the segmentation task is a new topic, where only a few attempt of exploration has been done. A first simple approach could be considering the depth map as a standard grayscale image, and segment it with standard image segmentation techniques. However, a better way of considering the geometric information comes straightforward from the classical 3D reconstruction problems.

The proposed algorithm requires the three dimensional position of each sample in the scene. To compute this information the depth value corresponding to each pixel in the color image is necessary together with the calibration parameters. Many passive and active methods (e.g. laser scanners or ToF sensors) directly provide depth information while in order to get depth information from the standard output of a stereo vision system, a further step is necessary. In fact, the output of a standard stereo vision algorithm is a disparity image $\tilde{D}_S$, that for each point $\mathbf{p}_i$ represents the distance in image space between its position in the left image of the stereo pair I_L, and its position in the right image I_R. Given the focal length f of the rectified cameras, and the baseline b of the stereo pair, it is possible to obtain a depth-map for the stereo pair D_S from the disparity map $\tilde{D}_S$ by applying the well-known equation:

$$D_S = \frac{bf}{\tilde{D}_S}. \tag{2}$$

Given the depth-map D_S, it is now possible to obtain a description of the geometrical information, obtaining for each point $\mathbf{p}_i \in \mathcal{S}, i = 1, ..., N$ the geometric information vector:

$$\mathbf{p_i^g} = \begin{bmatrix} x(\mathbf{p}_i) \\ y(\mathbf{p}_i) \\ z(\mathbf{p}_i) \end{bmatrix} \tag{3}$$

After getting the depth map D_T we can compute the three dimensional locations corresponding to the image samples: for each point $\mathbf{p}_i \in \mathcal{S}$, given its depth-map value $D_T(\mathbf{p}_i)$, its coordinates in the depth image $[u(\mathbf{p}_i), v(\mathbf{p}_i)]$, and the intrinsic camera matrix K_T, it is possible to obtain its 3D coordinates by simply inverting the projection equation:

$$\begin{bmatrix} x(\mathbf{p}_i) \\ y(\mathbf{p}_i) \\ z(\mathbf{p}_i) \end{bmatrix} = D_T(\mathbf{p}_i) K_T^{-1} \begin{bmatrix} u(\mathbf{p}_i) \\ v(\mathbf{p}_i) \\ 1 \end{bmatrix} \tag{4}$$

Given this equation, it is immediate to understand how for each point $\mathbf{p} \in \mathcal{S}$, the 3-D vector that accounts for the geometrical information is:

$$\mathbf{p_i^g} = \begin{bmatrix} x(\mathbf{p}_i) \\ y(\mathbf{p}_i) \\ z(\mathbf{p}_i) \end{bmatrix} \tag{5}$$

2.3. Combined color and depth segmentation

In order to better balance the relevance of the two kinds of information in the merging process, both the color information vectors $\mathbf{p}_i^c, i = 1, ..., N$ and the geometric information vectors $\mathbf{p}_i^g, i = 1, ..., N$ are normalized by the variance of the L and the z components respectively:

$$\sigma_L^2 = var\{L(\mathbf{p}_i)\}, i = 1, ..., N$$

$$\sigma_z^2 = var\{z(\mathbf{p}_i)\}, i = 1, ..., N$$

$$\begin{bmatrix} \bar{L}(\mathbf{p}_i) \\ \bar{a}(\mathbf{p}_i) \\ \bar{b}(\mathbf{p}_i) \end{bmatrix} = \frac{1}{\sigma_L} \begin{bmatrix} L(\mathbf{pi}_i) \\ a(\mathbf{p}_i) \\ b(\mathbf{p}_i) \end{bmatrix}$$

$$\begin{bmatrix} \bar{x}(\mathbf{p}_i) \\ \bar{y}(\mathbf{p}_i) \\ \bar{z}(\mathbf{p}_i) \end{bmatrix} = \frac{1}{\sigma_z} \begin{bmatrix} x(\mathbf{pi}_i) \\ y(\mathbf{p}_i) \\ z(\mathbf{p}_i) \end{bmatrix}$$

From these normalized color and geometric information vectors, it is possible to obtain the set of joint features vectors $\mathbf{p}_i^f, i = 1, ..., N$ according to the formula:

$$\mathbf{p}_i^f = \begin{bmatrix} \bar{L}(\mathbf{p}_i) \\ \bar{a}(\mathbf{p}_i) \\ \bar{b}(\mathbf{p}_i) \\ \lambda\bar{x}(\mathbf{p}_i) \\ \lambda\bar{y}(\mathbf{p}_i) \\ \lambda\bar{z}(\mathbf{p}_i) \end{bmatrix}, i = 1, ..., N \quad (6)$$

where λ is a parameter that allows the tuning of the contribution of the color and the geometric information.

The set of joint features vectors $\mathbf{p}_i^f, i = 1, ..., N$ is a set of characteristic features that allows for the synergic interaction of color and geometric information for the segmentation task, that is obtained in a very natural way, well reflecting the real scene distribution in both color and geometry. Given the set of vectors $\mathbf{p}_i^f, i = 1, ..., N$, a direct way to perform segmentation is the application of a clustering algorithm to this set. Numerous clustering techniques are currently available, more or less accurate and more or less time expensive. We decided to apply the standard k-means clustering algorithm because of its convergence speed. We also introduced a final refinement step that consists in removing regions smaller than a threshold t not connected to the main region of each cluster.

Finally note how the proposed segmentation algorithm need only three parameters in order to provide effective scene segmentation results:

- the parameter λ, that allows for tuning the distribution of decisional power between the color and the geometric information
- the parameter k, i.e., the number of segments that are produced by the segmentation algorithm (it is the number of clusters k of the k-means algorithm).
- (Eventually) the parameter t, i.e., the size of the smallest allowed unconnected region.

3. Experimental Results

To evaluate the performance of the proposed algorithm we tested it over some video sequences from standard multiview plus depth datasets. Figure 1 shows the performance of the proposed algorithm on the *breakdancers* sequence from Microsoft Research [8]. It refers to the first frame of view 0 and shows the results of the segmentation using only color, only geometry and both of them (the parameters used in this example are $\lambda = 5$ and $k = 15$). In particular note how by using only color information (Figure 1c) the noise in the camera image and some complex color patterns lead to an oversegmentation of the image and to some inconsistent region patterns, e.g. on the back of the breakdancer. Depth information is less noisy and leads to better performances (Figure 1d) but it is not able to capture some objects, e.g. the two spectators on the right that are very close to the wall and do not have a significant depth difference from the background. Figure 1e shows the results with the proposed joint segmentation algorithm: it is possible to see that it is able to locate all the objects in the scene, both the ones that have a significant depth difference and the ones that can be located only from the color data. It is also less noisy than the one based on color information.

Figure 2 instead shows the performance of the algorithm on a frame from the *orbi* sequence. Again notice how by using together color and depth information (Figure 2e) it is possible to recognize the foreground objects but also the board on the back that is not represented in depth data. By comparing Figures 2e and 2f it is also possible to see how the segmented image is modified by the final step that removes unconnected regions.

4. Application of the proposed scheme in depth compression

3D video and depth image-based (DIBR) representations require the transmission of one or more views together with the corresponding depth maps. While color data can be compressed with standard video compression techniques [4], ad-hoc solutions for depth data allows to obtain better performances than standard image or video compression algorithms. Depth maps are usually made by smooth regions divided by sharp edges and a common solution employed in ad-hoc compression solutions for depth data consist in locating the edges and objects in the scene through an initial segmentation step and then applying some ad-hoc

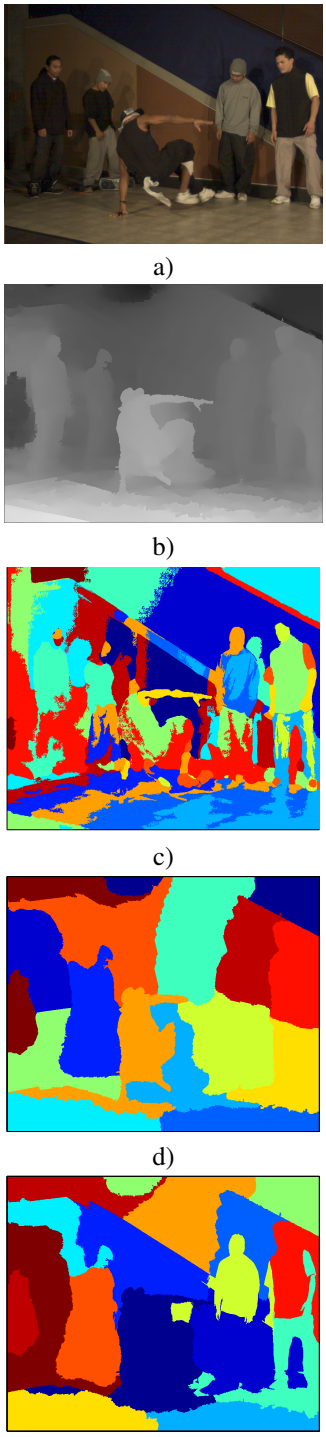

Figure 1. Segmentation of frame 0 (view 0) of the *breakdancers* sequence: a) color image; b) corresponding depth map; c) segmentation on the basis of color information only; d) segmentation on the basis of depth information only; e) Proposed joint segmentation scheme.

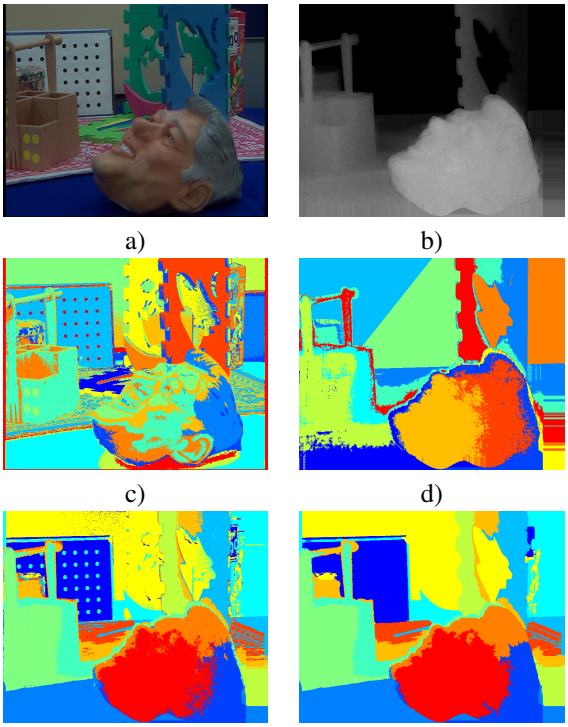

Figure 2. Segmentation of frame 0 of the *orbi* sequence: a) color image; b) corresponding depth map; c) segmentation on the basis of color information only; d) segmentation on the basis of depth information only; e) Proposed joint segmentation scheme before the final refinement; f) Proposed scheme with the further refinement stage.

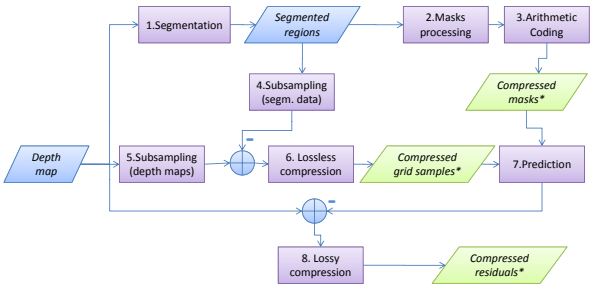

Figure 3. Architecture of the depth compression scheme of [7]

compression methods [7, 5]. Considering that color information is usually available together with the depth data, the proposed segmentation algorithm can be exploited to improve the performance of these methods. In this connection we considered the compression scheme of [7] and replaced the initial segmentation step with the algorithm proposed in this paper. In [7] the segmentation step is used in order to identify and extract the edges and the main objects in the scene. Segmented data is then compressed. In the subsequent step an ad-hoc algorithm is used to predict the surface shape from the segmented regions and a set of reg-

ularly spaced samples. Finally the prediction residuals are compressed using standard image compression techniques. Figure 3 shows the architecture of this compression algorithm, while a complete description of the system can be found in [7]. By exploiting also the color information the proposed algorithm is able to perform a more accurate segmentation of the depth map thus allowing better compression performance. Figure 4 refers to a preliminary test on the first frame of view 1 of the *breakdancers* sequence and shows a comparison between the results presented in [7], the ones of the same scheme with the proposed joint segmentation scheme and JPEG2000 (note that the scheme of [7] was targeted to high bitrate or near lossless compression). The proposed scheme allows to obtain a small but noticeable gain (on average 0.5 dB, but up to 1 dB at around 0.12 bpp). The gain is limited by the fact that here the target is just depth compression alone and so depth segmentation alone is already quite satisfactory, but the proposed scheme is particularly suited to be used in joint depth and color compression schemes where the redundancy between the two kind of data must be exploited and the proposed segmentation scheme is a very good solution to locate consistent regions in the two scene descriptions. This problem will be the subject of future research.

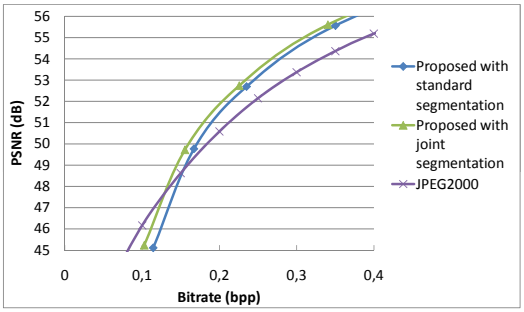

Figure 4. Depth compression performance on the breakdancers sequence.

5. Conclusions

In this paper a novel scene segmentation framework that accounts for both color and geometry is presented. The goal of the paper is to give a representation of both kind of information in a unified way, and to provide a solid background for the application of various clustering techniques. In the preliminary work presented in this paper we introduced a unified vector representation for the samples including color and depth data and applied the k-means clustering algorithm. Experimental results show the effectiveness of the proposed scene segmentation method. The choice of the optimal clustering technique is far beyond the purposes of this paper, and will be the subject of future investigation. We also presented one particular application of the developed scene segmentation algorithm in depth compression for DIBR schemes. In this direction another aspect that will be investigated is the exploitation of the proposed scheme in joint depth and color compression. The variety of possible applications is anyway not limited to the data compression problem, and it is in continuous expansion.

References

[1] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603 –619, may. 2002. 1

[2] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2):167–181, 2004. 1

[3] A. Frick, F. Kellner, B. Bartczak, and R. Koch. Generation of 3d-tv ldv-content with time-of-flight camera. In *Proc. of 3DTV Conf.*, 2009. 1

[4] ISO/IEC MPEG & ITU-T VCEG. Joint Draft 8.0 on Multiview Video Coding, Jul. 2008. 3

[5] S. Milani and G. Calvagno. A depth image coder based on progressive silhouettes. *Signal Processing Letters, IEEE*, 17(8):711 –714, aug. 2010. 4

[6] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42, 2001. 1

[7] P. Zanuttigh and G.M. Cortelazzo. Compression of depth information for 3d rendering. In *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009*, pages 1 –4, may. 2009. 4, 5

[8] Lawrence C. Zitnick, Sing B. Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM Trans. Graph.*, 23(3):600–608, 2004. 3