ACCURATE DEPTH-COLOR SCENE MODELING FOR 3D CONTENTS GENERATION WITH LOW COST DEPTH CAMERAS

Massimo Camplani, Tomás Mantecón and Luis Salgado

ABSTRACT

In this paper, we present a depth-color scene modeling strategy for indoors 3D contents generation. It combines depth and visual information provided by a low-cost active depth camera to improve the accuracy of the acquired depth maps considering the different dynamic nature of the scene elements. Accurate depth and color models of the scene background are iteratively built, and used to detect moving elements in the scene. The acquired depth data is continuously processed with an innovative joint-bilateral filter that efficiently combines depth and visual information thanks to the analysis of an edge-uncertainty map and the detected foreground regions. The main advantages of the proposed approach are: removing depth maps spatial noise and temporal random fluctuations; refining depth data at object boundaries, generating iteratively a robust depth and color background model and an accurate moving object silhouette.

Index Terms— 3D content generation, depth map filtering, bilateral filter, mixture of Gaussians, active depth cameras, Kinect.

1. INTRODUCTION

In the last years growing attention has been paid to 3D contents generation for multimedia applications to augment user experience with the interaction of virtual objects and scenarios. In particular, the great success of low-cost depth cameras with good resolution and acquisition frame rate, such as [1], is multiplying the number of 3D-based applications proposed everyday by developers and researchers. Initially proposed for Human Computer Interface applications, such as gaming [2], low-cost depth cameras have been successfully employed in other research areas such as robot-navigation [3], 3D scene segmentation [4] and object recognition [5].

However, low cost depth cameras's data is in general affected by different types and levels of noise (see Section 2); even if this issues do not critically compromise the performance of applications such as gaming, it is fundamental to develop efficient strategies to improve the depth data accuracy and hence to broaden the application possibilities of low-cost depth cameras. Several techniques have been developed to improve depth maps accuracy, that aim at removing artifacts, refining object boundaries, and generating a smooth depth map. Simple smoothing approaches lead to poor results since the filtered depth data appear excessively blurred near to depth discontinuities; hence, edge preserving filtering techniques, such as the bilateral filter [6], have to be deployed to avoid this effect. The weights of the bilateral filter are selected as a function of a photometric similarity measure of the neighbor pixels; non-similar neighbor pixels are excluded from the filtering process and the blurring effect is reduced.

This work has been supported by the Ministerio de Economía y Competitividad of the Spanish Gov. under Proj. No. TEC2010-20412 (Enhanced 3DTV). Massimo Camplani would like to acknowledge the EU and UPM for supporting his activities through the Marie Curie-Cofund research grant.

This idea has been extended in the joint-bilateral filter [7] in which the weights are selected as a function of the properties of another *guidance* image. As far as the low-cost cameras' depth maps is concerned, standard bilateral filter have been applied to improve depth maps of the segmentation dataset proposed in [4]; the depth maps used in the object recognition dataset [5] have been processed with a recursive median filter to restore invalid depth measurements. The approach proposed in [8] is based on the analysis of motion vectors to register images in time, using a non-causal spatio-temporal median filtering. Good qualitative results are reported for some image areas, but this strategy does not tackle the noisy boundaries problem; moreover computational requirements are another important drawback of this approach. In [9] a GPU-based filtering system is presented: normalized convolution is used for the restoration of invalid depth measurements, followed by an edge-preserving guided filter. Although operating in real-time, limited depth map improvement is obtained mainly because only depth information is considered by the filters: erroneous depth values are interpolated for invalid depth measurements, and the objects noisy borders in depth are preserved and eventually blurred. In [10] an adaptive joint-bilateral filtering system fuses depth, visual and temporal information to improve the accuracy of the depth; its main drawback is that it cannot be directly adapted to scenes containing moving objects.

We propose in this paper an accurate depth-color scene modeling approach for 3D contents generation, based on some of the concepts developed in [10]. It allows improving the accuracy of the depth maps for both, static and moving objects in indoor environments considering their different dynamic nature. Accurate depth-based and color-based models of the scene background are iteratively built, and they are used to detect moving objects in the scene thus allowing efficiently filtering and refining their corresponding depth maps. The acquired depth data is continuously processed with an innovative joint-bilateral filter that efficiently combines depth and visual information to improve the depth map accuracy of both, static and moving objects, by considering the concept of an edge-uncertainty map. The innovative features are an efficient filtering process based on the introduction of the edge-uncertainty map and the extension of this modeling/filtering framework to scenarios that include also moving objects; in this way the proposed system results very suitable for the generation of 3D-content of dynamic indoor scenes.

2. DEPTH MAP ERRORS

Depth data provided by low-cost active depth cameras, such as [1], is affected by different type of errors. As it can be noticed in Figure 3(b), errors result in noisy object boundaries, incoherent pixel neighborhood depth values in flat regions such as walls, and pixels for which the depth measurements cannot be estimated (*nmd* pixels) that form non measured regions (marked in red in the figure). Moreover, depth measurements relative to a static object present random

temporal fluctuations. For more details, about depth camera noise see [10]. The proposed approach aims at reducing the effect of these errors in the depth map to accurately generate 3D contents.

3. SYSTEM OVERVIEW

The proposed depth-color modeling system is constituted by three main blocks as reported in Figure 1: a color-depth segmentation module detects moving objects in the scene and discriminate between foreground (Fg) and background (Bg) pixels; the acquired depth maps (D) are filtered with an innovative joint-bilateral filter by considering the estimated color (I_m) and depth (D_m) model and the detected Bg and Fg regions; in the third module the depth model and color model of the scene are iteratively updated.

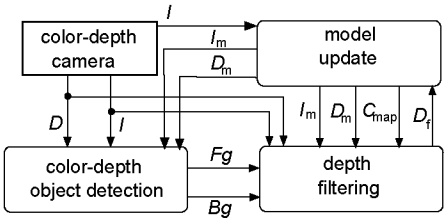

Fig. 1. Block diagram of proposed system.

3.1. Object detection and Model Update

The proposed strategy is based on the continuous estimation of a color-based model (I_m) and a depth-based model (D_m) of the static (background) objects; the model is used to improve depth map filtering and to accurately detect moving objects in the scene. In our approach the Mixture of Gaussian background modeling algorithm [11] is proposed: it is a popular algorithm that allows to accurately estimating quasi-static backgrounds, adapting to new background configurations and gradual changes; each pixel is modeled independently as a mixture of Gaussian distributions. During the object detection phase a statistical test is performed to detect those pixels that do not belong to the MoG distribution and are marked as foreground pixels. In particular, Fg and background Bg masks are estimated considering the two models independently; this information is used to properly adapt the filter characteristics to Bg and Fg pixels. It is worth noting that are considered *real* Fg pixels only those classified as Fg using D_m (depth-Fg). The pixels identified as Fg using I_m (color-Fg) are used to classify *nmd* pixels (there is no depth information associated to them), and to refine the boundaries of the depth-Fg regions. The depth model D_m is updated with the filtered depth maps D_f; thus allowing to reduce temporal random fluctuations of depth measurements for Bg pixels and to build the consistency map C_{map} employed to interpolate the *nmd* pixels. C_{map} indicates the reliability of depth measurements in D_m: the greater its value the higher its reliability; it is computed as the occurrence number of depth measurements that belong to the background model. I_m is updated with the color information I.

3.2. Data Filtering

The proposed filtering strategy is based on an innovative joint-bilateral filter that reduces the spatial noise of the acquired depth map while refining object boundaries and interpolating coherent depth values for the *nmd* pixels. The block diagram of the filter is shown in Figure 2. The filters used for Bg and Fg pixels are different, for the lack of space we report here only the filter equations for the Bg case and highlight the differences with the Fg case.

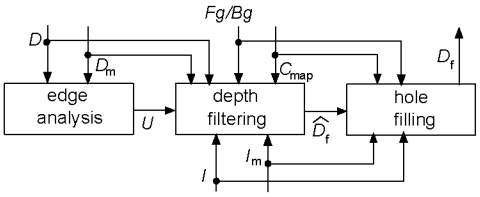

Fig. 2. Block diagram of depth filtering module.

Let us define D^p the pixel in the depth map D at the position p, and D^p_m the corresponding pixel in the depth-based model D_m. The depth value obtained with the joint-bilateral filter is:

$$\widehat{D}^p_f = \frac{1}{k^p} \sum_{q \in \Omega^p} D^q f(p,q) \widetilde{g}\left(I, D, I_m, D_m, U\right) \qquad (1)$$

where Ω^p is the neighborhood of the pixel at position p, $f(\bullet)$ is a smoothing Gaussian function known as the *spatial term* of the joint-bilateral filter, and the function $\widetilde{g}(\bullet)$ is the *range term* of the joint-bilateral filter that determines the weights of $f(\bullet)$ by measuring the pixels similarity. It is worth noting that sub-indexes p and q are not included in the range term in equation 1 for readability. U is the edge-uncertainty map and its values determine which model (I_m or D_m) has to be used to estimate the pixels similarity. As previously mentioned, the depth map presents noisy object boundaries with non reliable depth data; for this reason we propose to filter boundaries' pixels considering the color information. In fact, discontinuities in the visual domain in the neighborhood of the depth discontinuity help to consider in the filter only pixels that likely belong to the same object, thus refining the depth map at object boundaries. On the contrary, to obtain a smoothing effect in homogeneous depth regions only depth data is considered. The edge-uncertainty map U, is a binary mask that identifies regions in surrounding depth discontinuities in D that are characterized by a low correlation in the color domain; it is obtained by analyzing the gradient of D and the color data I. The range term $\widetilde{g}(\bullet)$ is thus evaluated as:

$$\widetilde{g}\left(\bullet\right) = \omega_1 g_1\left(\|D^p - D^q_m\|\right) + \omega_2 g_2\left(\|I^p - I^q_m\|\right) \qquad (2)$$

where $g_1(\bullet)$ and $g_1(\bullet)$ are the range terms (Gaussian functions) that measure respectively the similarity for p pixel with respect to the q pixel of the depth and color models. The binary weights ω_1 and ω_2 depend on the values of U and are defined such that $\omega_1 = xor\left(U_p, U_q\right)$ and $\omega_2 = not\left(\omega_1\right)$. In particular, if the filtered pixel p or its neighbor q is in the uncertainty zone, the similarity range term is calculated considering I_m and ω_1 is set to 0. In case of foreground pixels $\widetilde{g}(\bullet)$ is calculated by analyzing the D instead of D_m, U identifies the regions near depth the discontinuities of D and I is used instead of I_m. Moreover, the boundaries of the depth-Fg regions are refined by evaluating the corresponding more accurate color-Fg regions boundaries.

The obtained filtered depth map $\widehat{D}_f$ is then processed to interpolate values for the *nmd* pixels using consistent depth map values in their neighborhood. These new values are obtained by applying the joint-bilateral filter shown in equation 3. This filter is applied to all the *nmd* pixels that have in their neighborhood a significant number of pixels with reliable depth value.

$$D^p_f = H\left(C_{map}, \Omega^p\right) / k^p \sum_{q \in \Omega^p} \widehat{D}^q f(p,q) g_2(\|I^p_m - I^q_m\|) \qquad (3)$$

where $H\left(C_{map}, \Omega^p\right)$ evaluates the reliability of the depth values in the neighborhood Ω^p (see [10] for more details); $f(\bullet)$ is the spatial term and $g_2(\bullet)$ is the range term that considers color information.

It is worth noting that equation 3 assumes that the *nmd* pixel at position p is part of the background; this classification is performed by considering the color-Fg mask. If the *nmd* pixel belongs to the color-Fg mask, equation 3 is applied with $H(\bullet)$ that does not consider the C_{map} of the model (all color-Fg pixels in the neighborhood are reliable) and $g_2(\bullet)$ uses the color image I.

4. RESULTS

Figure 3 reports the data coming from the low-cost active depth camera: visual information (a) and raw depth data (b), that presents several noise related problems (described in Section 2) such as: *nmd* pixels (marked in red), very noisy object boundaries and spatial noise. We have tested the proposed approach with three different datasets: two are composed by sequences acquired in our laboratories, containing respectively a static and a dynamic scene; the third one is the object dataset proposed in [5].

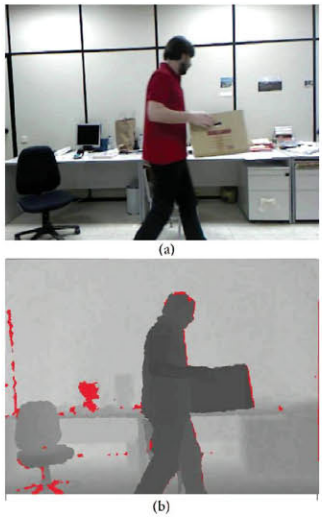

Fig. 3. Visual (a) and depth (b) obtained with the depth camera.

The static scene reported in Figure 4 has been used to test the performance of the strategy for different *color conditions* in depth-borders areas. In particular, we analyze the boundaries area of the two overlapped boxes: the brown one at the front, placed at 146 cm from the camera, and the white one behind on the left, placed at 186 cm from the camera. We test the proposed approach under four different *color conditions*: the first case, called HC, is characterized by a high contrast on the box border in the color domain (Figure 4(a)); in the second case, called NC, we employed the proposed algorithm without considerig the color information, thus simulating the case of an empty U map due to the presence of a depth discontinuity and a high correlation in the color domain; for the third and the fourth case (Figure 4(b) and (c)) we introduced colored patches to test the proposed approach in a border containing misaligned color and depth discontinuities, these case are called GC and WC. Finally we compare our approach with the algorithm proposed in [9] that is based on guided filtering (GF) considering only depth information. Figure 5 shows the performance of the proposed approach with respect to the different *color conditions*, the Normalized Absolute Error (NAE) is used to measure the algorithm accuracy. As expected, the NAE values obtained in the HC case (red line) are lower than the ones obtained in the NC case (blue line); textured color borders reduce the performance of the propoed strategy: in the GC case (green line) NAE values are increased since the presence of smaller *similar color* regions decreases the effective filtering area; the color-depth combination guarantees lower NAE values, in the case of misleading white patches, WC case (cyan line), with respect to the NC case. These tests demonstrate that the combination of color and depth data is fundamental to reduce the noise level present at object borders (black line) also in the cases in which color similarity and depth similarity do not correspond. It is worth noting that the proposed algorithm overperforms the depth based approach proposed in [9] (magenta line). The resulting depth map obtained in the HC and NC cases are reported in Figure 4 (d) and (e).

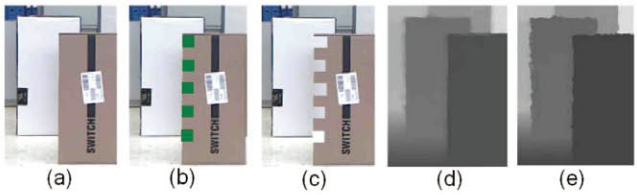

Fig. 4. Object boundaries tests with different color condition: HC (a), GC (b), WC (c), depth map obtained in the HC case (d), depth map obtained in the NC case (e).

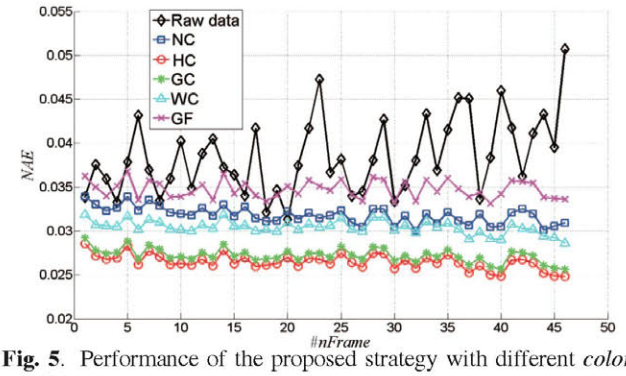

Fig. 5. Performance of the proposed strategy with different *color conditions*.

Qualitative results of the proposed strategy's performance applied to a static scene are reported in Figure 6, where details of sample images of dataset [5] are reported. We compare the proposed strategy with the algorithm proposed in [9], based on the combination of guided filtering and normalized convolution, and with the recursive median filter proposed in [5] to correct the *nmd* pixels. As it can be observed, the proposed method improves significantly the accuracy of the depth maps: boundaries are completed and refined, leading to locally smooth regions with accurate depth values for *nmd* pixels, thanks to the efficient combination of depth and color information. On the contrary, the other approaches, that do not use color data, do not guarantee an accurate depth map refinement at object borders where *nmd* pixels interpolated values tend to significantly over(under)-pass the actual object borders (i.e. the red cap).

Sequences taken in an indoor environment (Figure 3) have been used to test the performance of the proposed approach in case of scenes containing moving objects through the evaluation of the detection accuracy with respect to a ground truth containing the Fg and Bg masks; the sequence is composed by 150 frames. As a measure

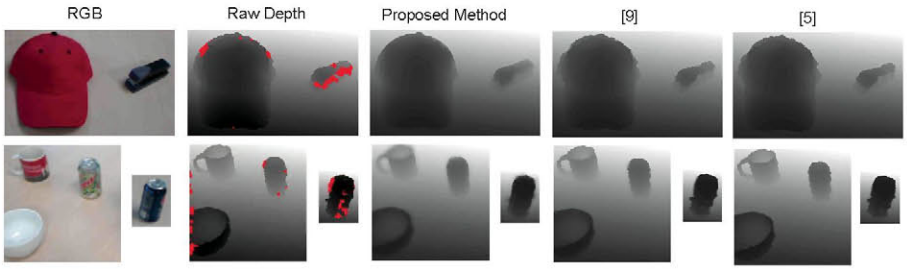

Fig. 6. Processed images from dataset [5]: color images (column 1), raw depth maps (column 2), proposed method results (column 3), [9] results (column 4), [5] results (column 5).

of algorithm performance, three parameters are used: False Positive (FP), that is the fraction of the Bg pixels that are marked as Fg; False Negative (FN), that is the fraction of Fg pixels that are marked as Bg, and the similarity measure S proposed in [12], that is a non-linear measure that fuses FP and FN indexes (close to 1 if detected Fg regions are similar to the real ones, close to 0 if they are different). Table 1 reports the results obtained using the proposed approach, and those obtained considering only D_m or the I_m. As it can be noticed, the single use of I_m leads to a high value of FN due mainly to the well known problem of color camouflage. On the contrary, by considering only D_m, very compact foreground regions are obtained thus leading to low FN; nevertheless, considering only depth information results in a higher value of FP, due to the noisy depth measurements at the object boundaries. The proposed combination of both models reduces simultaneously both FP and FN values, and guarantees a very high value of S. Figure 7(a) reports a detail of the refined silhouette of the moving object obtained with the proposed approach; as it can be noticed, the incorporation of the color-Fg mask improves the accuracy at object boundaries, fitting much better the actual moving object silhouette.

Model	FN	FP	S
Proposed: $D_m + I_m$	2.58%	0.54%	0.9
I_m	14%	2.1%	0.68
D_m	1.8%	6.7%	0.6

Table 1. Foreground/Background modeling accuracy.

Fig. 7. Filtered depth map: with color-Fg mask (a) and without the color-Fg mask (b).

5. CONCLUSION

In this paper, we have presented a depth-color scene modeling for 3D content generation of dynamic indoor environments. The proposed approach combines depth and visual data provided by a low-cost active depth camera to improve the depth map accuracy. The MoG algorithm is used to iteratively build an accurate color and depth model of the background elements in the scene and to identify foreground pixels. An innovative joint-bilateral filter is proposed that efficiently combines depth and visual information to improve the depth data map accuracy by considering an edge-uncertainty map. The filtered depth map is then used to iteratively build a coherent and reliable depth-based model of the scene. Results demonstrate that the proposed system dramatically improves the accuracy of the 3D contents provided by the low-cost active depth camera.

6. REFERENCES

[1] "Microsoft Corporporation. Kinect for Xbox 360," .

[2] A. Bleiweiss et al., "Enhanced interactive gaming by blending full-body tracking and gesture animation," in *ACM SIGGRAPH ASIA 2010 Sketches*, 2010, pp. 34:1–34:2.

[3] P. Benavidez and M. Jamshidi, "Mobile robot navigation and target tracking system," in *System of Systems Engineering (SoSE), Int. Conf. on*, 2011, pp. 299–304.

[4] N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," in *Computer Vision (ICCV Workshops), IEEE Int. Conf. on*, 2011, pp. 601 –608.

[5] Kevin Lai, Liefeng Bo, and Xiaofeng Ren, "A large-scale hierarchical multi-view rgb-d object dataset," *Robotics and Automation (ICRA)*, pp. 1817–1824, 2011.

[6] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Computer Vision, Int. Conf. on*, 1998, pp. 839–846.

[7] G. Petschnigg et al., "Digital photography with flash and no-flash image pairs," in *ACM Transactions on Graphics (TOG)*, 2004, vol. 23, pp. 664–672.

[8] S. Matyunin et al., "Temporal filtering for depth maps generated by Kinect depth camera," in *2011 3DTV Conference*, 2011, pp. 1–4.

[9] J. Wasza et al., "Real-time preprocessing for dense 3-D range imaging on the GPU: Defect interpolation, bilateral temporal averaging and guided filtering," in *Computer Vision (ICCV Workshops), IEEE Int. Conf. on*, 2011, pp. 1221–1227.

[10] M. Camplani and L. Salgado, "Efficient spatio-temporal hole filling strategy for Kinect depth maps," in *SPIE Electroinc Imaging Conference*, 2012, vol. 8290, p. 82900E.

[11] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, IEEE Int. Conf. on*, 1999, pp. 246–252.

[12] L. Li et al., "Statistical modeling of complex backgrounds for foreground object detection," *Image Processing, IEEE Transactions on*, vol. 13, no. 11, pp. 1459 –1472, 2004.