3D Modeling from Multiple Images

Wei Zhang[1], Jian Yao[2], and Wai-Kuen Cham[1]

[1] Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
[2] College of Computer Science and Electronic Information, Guangxi University, Nanning 530004, China

Abstract. Although the visual perception of 3D shape from 2D images is a basic capability of human beings, it remains challenging to computers. Hence, one goal of vision research is to computationally understand and model the latent 3D scene from the captured images, and provide human-like visual system for machines. In this paper, we present a method that is capable of building a realistic 3D model for the latent scene from multiple images taken at different viewpoints. Specifically, the reconstruction proceeds in two steps. First, generate dense depth map for each input image by a Bayesian-based inference model. Second, build a complete 3D model for the latent scene by integrating all reliable 3D information embedded in the depth maps. Experiments are conducted to demonstrate the effectiveness of the proposed approach.

Keywords: 3D modeling, Depth map, Fusion.

1 Introduction

As a popular research topic, image-based 3D scene modeling has attracted much attention in the past decades. In short, the task is to build a realistic 3D representation for the latent scene from a collection of images. Such technique can be widely applied in various areas such as robot navigation, virtual reality, computer games and art.

In this paper, an algorithm is presented which is capable of creating a complete and detailed 3D model from multiple views. The reconstruction proceeds by a two-step process. Firstly, generate dense depth map for each view. Then, integrate all reliable 3D information embedded in these input views into a single model through patch-based fusion. In specific, a Bayesian-based framework is employed to infer the depth maps of the multiple input images. However, each depth map can only reveal the scene's 3D information at one viewpoint. For a large and complex scene, a single depth map is insufficient to produce the desirable detailed and complete structure. Therefore, a patch-based fusion scheme is adopted to integrate all individual modeling structures into a single one.

Besides, due to the influence of geometric occlusion, specular reflection and image noise, the resulting depth maps may contain some outlier pixels that have inaccurate depth estimates. Hence, it is necessary to introduce a refinement step to ensure that the tessellated surface patches at each view are derived only from reliable points and thus avoid fusing these outliers into the final 3D model.

L. Zhang, J. Kwok, and B.-L. Lu (Eds.): ISNN 2010, Part II, LNCS 6064, pp. 97–103, 2010.

The remainder of this paper is organized as follows. Section 2 reviews some related work. Section 3 introduces a Bayesian-based inference model for depth map generation. Section 4 describes how to build a complete 3D model by patch fusion. Experimental results are shown in Section 5. Section 6 gives some concluding remarks.

2 Related Work

Since lots of efforts such as [1,2,3,4,5,6,7,8,9] have been made to develop new approaches for modeling complex scene from a single or multiple images, we just refer some methods which are the most related to ours.

In this work, the depth map recovery problem is formulated in a improved Bayesian-based framework [3], which can be regarded as an extension of [4] and [5]. However, some new contributions have been done. For example, the hidden consistency variables are introduced to smooth and integrate the depth maps at the same time. A data-driven regularizer is adopted to preserve the discontinuities at the image boundaries. A new visibility prior is defined based on the transformation consistency between different depth maps, which is used to account for the occlusion and noise problem. Also, a bilateral consistency prior is developed to impose the spatial smoothness in one depth map and the temporal consistency among different depth maps. Moreover, the EM (Expectation Maximization) optimization is implemented in a coarse-to-fine resolution manner.

Narayanan et al. [6] presented a technique, *Virtualized Reality*, to build a complete surface model by merging depth maps into a common volumetric space. They designed a special system, 3D Dome which consists of 51 cameras, to capture images at multiple viewpoints. Also, conventional multi-baseline stereo technique was adopted to recover the dense depth maps. Goesele et al. [7] used a robust window-based matching method to produce depth map for each input image. The depth map result is not a dense one since only the pixels that can be matched with high confidence are reconstructed.

3 Depth Map Estimation

Given a collection of images taken from different viewpoints, the latent scene will be reconstructed under a Bayesian-based inference model, which is briefly described as follows. More details can be found in [3]. Finally, the latent scene will be represented by a set of depth maps.

From a small collection of N input images $\mathcal{I} = \{\mathcal{I}_i, i = 1, \ldots, N\}$ and a sparse set of N_p 3D scene points $\mathcal{Z} = \{Z_p, p = 1, \ldots, N_p\}$ precalculated based on camera self-calibration and stereo feature matching, we intend to estimate the unknown model $\boldsymbol{\theta} = (\mathcal{D}, \mathcal{I}^*)$ where $\mathcal{D} = \{\mathcal{D}_i, i = 1, \ldots, N\}$ and $\mathcal{I}^* = \{\mathcal{I}_i^*, i = 1, \ldots, N\}$ represent the sets of estimated depth maps and estimated images, respectively. In fact, $\mathcal{I}^*$ corresponds to the input image set $\mathcal{I}$. The variable τ represents the set of parameters that will be fixed or heuristically updated in our inference system. To efficiently deal with occlusion, specular reflection and image noise, we introduce a set of hidden

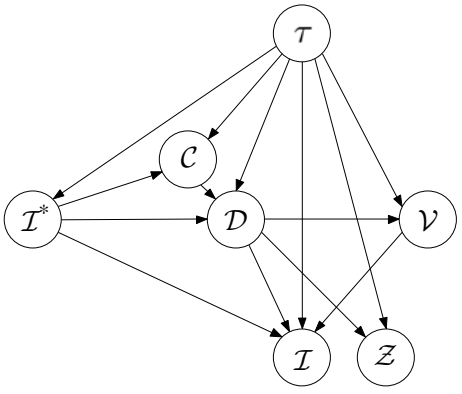

Fig. 1. Network representation of the joint probability decomposition. Arrows represent statistical dependencies between variables.

visibility variables $\mathcal{V} = \{\mathcal{V}_{j,\mathbf{x}_i} | \mathbf{x}_i \in \mathcal{I}_i, i, j = 1, \cdots, N\}$ based on priors of transformation consistencies in the geometrical sense where $\mathcal{V}_{j,\mathbf{x}_i}$ is a boolean variable that denotes whether the pixel $\mathbf{x}_i$ in $\mathcal{I}_i$ is visible or not in $\mathcal{I}_j$. In addition, a set of hidden consistency variables $\mathcal{C} = \{\mathcal{C}_{j,\mathbf{y}_i,\mathbf{x}_i} | \mathbf{x}_i \in \mathcal{I}_i, \mathbf{y}_i \in \mathcal{N}(\mathbf{x}_i), i, j = 1, \cdots, N\}$ are introduced to smooth and integrate the depth maps while ensuring consistencies among different depth maps and allowing discontinuities based on priors of local gradients of the estimated images. In specific, $\mathcal{C}_{j,\mathbf{y}_i,\mathbf{x}_i}$ is a boolean variable that denotes whether the pixels $\mathbf{x}_i$ and $\mathbf{y}_i$ are consistent or not via transformation w.r.t. $\mathcal{I}_j$. After defining all the variables $(\mathcal{I}, \mathcal{Z}, \mathcal{I}^*, \mathcal{D}, \mathcal{V}, \mathcal{C}, \tau)$, next step of the Bayesian modeling task is to choose a suitable decomposition of their joint probability $p(\mathcal{I}, \mathcal{Z}, \mathcal{I}^*, \mathcal{D}, \mathcal{V}, \mathcal{C}, \tau)$. The decomposition defines the statistical dependencies between the variables involved in our proposed model. Based on the proposed decomposition shown in Fig.1, the joint probability can be written as:

$$\begin{aligned} p(\mathcal{I}, \mathcal{Z}, \mathcal{I}^*, \mathcal{D}, \mathcal{V}, \mathcal{C}, \tau) = & p(\tau) p(\mathcal{I}^* | \tau) p(\mathcal{V} | \mathcal{D}, \tau) \\ & p(\mathcal{C} | \mathcal{I}^*, \tau) p(\mathcal{D} | \mathcal{I}^*, \mathcal{C}, \tau) \\ & p(\mathcal{Z} | \mathcal{D}, \tau) p(\mathcal{I} | \mathcal{I}^*, \mathcal{D}, \mathcal{V}, \tau). \end{aligned} \tag{1}$$

Each term of the decomposition in (1) will be introduced briefly as follows. $p(\tau)$ which defines the prior probability of all involved parameters is to assumed to be uniform and thus is ignored in this work. $p(\mathcal{I}^* | \tau)$ denotes the prior of the images to be estimated. In general, this term was introduced to enforce that the estimated images $\mathcal{I}^*$ look more like natural images. $p(\mathcal{V} | \mathcal{D}, \tau)$ is the consistent visibility prior that depends on $\mathcal{D}$ and τ. $p(\mathcal{C} | \mathcal{I}^*, \tau)$ is the bilateral consistency prior that depends on $\mathcal{I}^*$ and τ. $p(\mathcal{D} | \mathcal{I}^*, \mathcal{C}, \tau)$ is the prior on depth maps given $\mathcal{I}^*$,$\mathcal{C}$ and τ. $p(\mathcal{Z} | \mathcal{D}, \tau)$ is the likelihood of the input 3D scene points with known visibility values. It measures the similarity between the model and the input scene points and is used to preserve the correspondences appear in these precalculated 3D scene points. $p(\mathcal{I} | \mathcal{I}^*, \mathcal{D}, \mathcal{V}, \tau)$ is the likelihood of the input images,

which measures the similarity between the unknown model $\boldsymbol{\theta} = (\mathcal{D}, \mathcal{I}^*)$ and the input image data. In summary, the Bayesian-based inference problem can be recasted to estimate $\boldsymbol{\theta} = (\mathcal{D}, \mathcal{I}^*)$ as below:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{I}, \mathcal{Z}, \tau) = \arg\max_{\boldsymbol{\theta}} \int_{\mathcal{C}} \int_{\mathcal{V}} p(\mathcal{I}, \mathcal{Z}, \mathcal{I}^*, \mathcal{D}, \mathcal{V}, \mathcal{C}, \tau) d\mathcal{V} d\mathcal{C}. \quad (2)$$

In the implementation, the EM optimization strategy is adopted to solve (2) and produce the desired depth maps. Particularly, EM is implemented efficiently with a coarse-to-fine resolution scheme.

4 Create 3D Model by Patch Fusion

When the depth map is fixed, 3D structure of each image can be created by triangulation. Next, to seek a more complete and detailed model for the latent scene, we integrate these tessellated structures obtained at different views into a single model. However, although the above Bayesian-based inference model provides a fairly reliable way for depth map estimation, it is inevitable that some pixels may have inaccurate depth estimates due to the influence of geometric occlusion, specular reflection and image noise.

To remove the influence of these outlier pixels, depth map will be firstly refined with the guidance of pixel's visibility. A binary mask $\mathcal{M}_i$ $(i = 1, \ldots, N)$ is defined for each image $\mathcal{I}_i$ where $\mathcal{M}_i(\mathbf{x}_i) = 1$ denotes the depth estimate $d_i(\mathbf{x}_i)$ of pixel $\mathbf{x}_i$ in image $\mathcal{I}_i$ is reliable. Otherwise, it is unreliable. As a typical multi-view stereo method, our Bayesian-based inference system can impose effective constraints when points of the scene are visible in three views or more. Therefore, a criterion can be defined based on the visibility map as (3). If a pixel $\mathbf{x}_i$ in image $\mathcal{I}_i$ is visible at least k neighbor views $(k \geq 3)$, its depth estimate is regarded as reliable.

$$\mathcal{M}_i(x_i) = \begin{cases} 1 & \sum_{j=1}^{N} \mathcal{V}_{j,\mathbf{x}_i} \geq k, \\ 0 & otherwise. \end{cases} \quad (3)$$

As addressed in [10], the visibility map can be estimated in a straightforward way. While in this work, it is formulated into the Bayesian-based inference model by introducing a visibility prior $p(\mathcal{V}|\mathcal{D}, \tau)$ as mentioned in the last section. Hence, the visibility map of each image will be produced more robustly as a by-product of the above depth map estimation system.

However, since the visibility estimates may also contain outliers, an additional refinement step is introduced based on the criterion that: if the neighbors of a pixel have reliable depth estimates, this pixel should also have reliable depth estimate. Otherwise, the current depth estimate is probably unreliable. Since the outliers look like salt and pepper noise in the binary mask $\mathcal{M}_i$, an adaptive median filter is employed to remove them. The reasons of using median filter are as follows. Firstly, the visibility mask is a binary image, the value of each pixel can only be 1 or 0. Median filter use a neighborhood value to substitute the false one, so the filtered mask remains binary. Secondly, as a non-linear filtering technique, it works particularly well in removing shot and isolated noise with edge-preserving property.

After fixing the mask for each input image, we are able to preserve pixels with the reliable depth estimates and discard the outlier ones. For each image, a set of surface patches will be created by tessellating these points that have reliable depth estimates.

Motivated by the work on range image data [11,12], we adopt the volumetric fusion technique to integrate all the structure patches into a single 3D model due to some of its desirable properties such as resilience to noise, simplicity of design, and non-iterative operation. As in [11], a weighting function $\mathcal{W}(\mathbf{p})$ and a cumulative signed distance function $Dis(\mathbf{p})$ are defined in (4) and (5), respectively. $\mathbf{p}$ denotes a point of the structure. $Dis(\mathbf{p})$ is constructed by combining the signed distance functions $d_1(\mathbf{p}), \ldots, d_n(\mathbf{p})$ with their corresponding weight factors $w_1(\mathbf{p}), \ldots, w_n(\mathbf{p})$.

$$\mathcal{W}(\mathbf{p}) = \sum_{i=1}^{n} w_i(\mathbf{p}). \tag{4}$$

$$Dis(\mathbf{p}) = \frac{\sum_{i=1}^{n} w_i(\mathbf{p}) d_i(\mathbf{p})}{W(\mathbf{p})}. \tag{5}$$

In the implementation, the functions are casted in discrete voxel grid of a 3D volume. Finally, an isosurface corresponding to $Dis(\mathbf{p}) = 0$ can be extracted by employing Marching Cubes [13]. Please refer to [11] for more technical details.

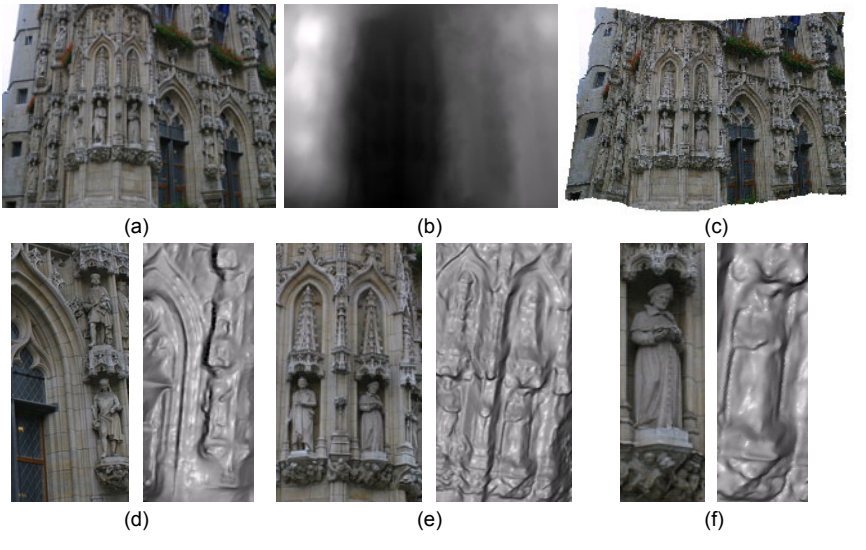

Fig. 2. Testing on *Cityhall* sequence. (a) shows one sample image of the sequence. (b) is the estimated depth map of (a). (c) shows the textured 3D modeling result. (d), (e) and (f) show some close-up comparisons between the fused untextured 3D model and the corresponding image view. In each pair, left shows a sight in the image, right shows the output 3D structure.

5 Experimental Results

In this section, the proposed algorithm are tested on different kinds of image sequences to demonstrate its effectiveness.

Cityhall shows a complex scene with significant depth discontinuities. 7 images are captured arbitrarily in a wide-baseline condition. Fig.2(a) shows one sample image. Fig.2(b) and (c) show the corresponding depth map and textured 3D model respectively. Some parts of the final fusion model are enlarged and compared with the image as shown in Fig.2(d), (e) and (f). Apparently, the proposed method produced a good 3D model with abundant details.

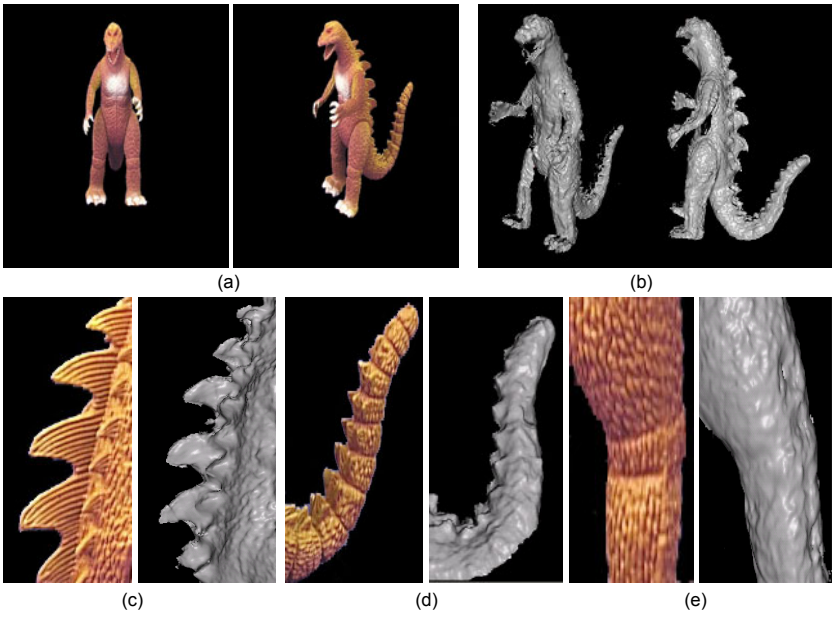

Fig. 3. Testing on *Dinosaur* sequence. (a) shows two sample images of the sequence. (b) shows two views of the fused complete 3D model (untextured). (c), (d) and (e) show some close-up comparisons between the fused untextured 3D model and the corresponding image view. In each pair, left shows a sight in the image, right shows the output 3D structure.

Dinosaur is a Turn-Table sequence which consists of 36 images [14]. This data is used to demonstrate that our method is able to build a complete and detailed 3D model. Two sample images are shown in Fig.3(a). Fig.3(b) shows two shots of the output 3D model reconstructed by fusing 36 structures. As shown in the close-up comparisons in Fig.3(c), (d) and (e), the generated 3D structure is highly faithful to the truth.

6 Conclusions

In this paper, we presented an image-based modeling method to create a realistic 3D model for complex scene. The motivation of this work is as follows. Each image reveals

a certain characteristic of the latent scene at one viewpoint. Hence, we intend to exploit the individual 3D information at each view and then combine the reliable estimates to produce a complete and more detailed 3D model for the latent scene. Experimental results demonstrated the effectiveness of our method. A complete 3D model can be built if enough images which contain all information about the latent scene are given. However, the proposed approach shares the common limitation of most 3D modeling methods. For example, it cannot work well when serious geometric occlusion, specular reflection or noise occurs in the input image sequence. In the future, we would like to fuse the input images to texture the generated 3D model.

References

1. Saxena, A., Sun, M., Ng, A.Y.: Make3D: Learning 3D scene structure from a single still image. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 31, 824–840 (2009)
2. Fitzgibbon, A., Wexler, Y., Zisserman, A.: Image-based rendering using image-based priors. In: Proceedings of ICCV, vol. 2, pp. 1176–1183 (2003)
3. Yao, J., Cham, W.K.: Consistent 3D modeling from multiple widely separated images. In: Proceedings of ECCV Workshop on WRUPKV. LNCS. Springer, Heidelberg (2006)
4. Strecha, C., Fransens, R., Gool, L.V.: Wide-baseline stereo from multiple views: a probabilistic account. In: Proceedings of CVPR, vol. 1, pp. 552–559 (2004)
5. Gargallo, P., Sturm, P.: Bayesian 3D modeling from images using multiple depth maps. In: Proceedings of CVPR, vol. 2, pp. 885–891 (2005)
6. Narayanan, P., Rander, P., Kanade, T.: Constructing virtual worlds using dense stereo. In: Proceedings of ICCV, pp. 3–10 (1998)
7. Goesele, M., Curless, B., Seitz, S.M.: Multi-view stereo revisited. In: Proceedings of CVPR, vol. 3, pp. 1278–1285 (2006)
8. Tan, P., Zeng, G., Wang, J., Kang, S., Quan, L.: Image-based tree modeling. In: Proceedings of SIGGRAPH, vol. 26(3) (2007)
9. Sinha, S.N., Steedly, D., Szeliski, R., Agrawala, M., Pollefeys, M.: Interactive 3D architectural modeling from unordered photo collections. ACM Trans. on Graphics (TOG) 27, 1–10 (2008)
10. Szeliski, R.: A multi-view approach to motion and stereo. In: Proceedings of CVPR, vol. 1, pp. 23–25 (1999)
11. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of SIGGRAPH, pp. 303–312 (1996), `http://grail.cs.washington.edu/software-data/vrip/`
12. Hilton, A., Illingworth, J.: Geometric fusion for a hand-held 3D sensor. Machine Vision and Applications 12(1), 44–51 (2000)
13. Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3D surface construction algorithm. In: Proceedings of SIGGRAPH, vol. 21, pp. 163–169 (1987)
14. Fitzgibbon, A.W., Cross, G., Zisserman, A.: Automatic 3D model construction for turn-table sequences. In: Koch, R., Van Gool, L. (eds.) SMILE 1998. LNCS, vol. 1506, pp. 155–170. Springer, Heidelberg (1998)