ISPRS Journal of Photogrammetry and Remote Sensing 139 (2018) 133–145

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Planarity constrained multi-view depth map reconstruction for urban scenes

Yaolin Hou [a], Jianwei Peng [a], Zhihua Hu [a], Pengjie Tao [a], Jie Shan [a,b,*]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[b] Lyles School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

Article history:
Received 10 September 2017
Received in revised form 3 January 2018
Accepted 1 March 2018

Keywords:
Planarity constraint
Multi-view depth map
Optimization
Segmentation
PatchMatch

ABSTRACT

Multi-view depth map reconstruction is regarded as a suitable approach for 3D generation of large-scale scenes due to its flexibility and scalability. However, there are challenges when this technique is applied to urban scenes where apparent man-made regular shapes may present. To address this need, this paper proposes a planarity constrained multi-view depth (PMVD) map reconstruction method. Starting with image segmentation and feature matching for each input image, the main procedure is iterative optimization under the constraints of planar geometry and smoothness. A set of candidate local planes are first generated by an extended PatchMatch method. The image matching costs are then computed and aggregated by an adaptive-manifold filter (AMF), whereby the smoothness constraint is applied to adjacent pixels through belief propagation. Finally, multiple criteria are used to eliminate image matching outliers. (Vertical) aerial images, oblique (aerial) images and ground images are used for qualitative and quantitative evaluations. The experiments demonstrated that the PMVD outperforms the popular multi-view depth map reconstruction with an accuracy two times better for the aerial datasets and achieves an outcome comparable to the state-of-the-art for ground images. As expected, PMVD is able to preserve the planarity for piecewise flat structures in urban scenes and restore the edges in depth discontinuous areas.

1. Introduction

Multi-view stereo, a popular image-based 3D reconstruction method, exploits several images taken from different views and applies dense matching techniques to produce a high-quality 3D representation of the scene. Such techniques can be divided into four groups (Furukawa and Ponce, 2010): (1) voxel-based (Faugeras and Keriven, 1998; Heise et al., 2015); (2) deformable polygonal-meshes-based (Esteban and Schmitt, 2004; Furukawa and Ponce, 2009); (3) depth-maps-based (Kolmogorov and Zabih, 2001; Goesele et al., 2006; Kwon et al., 2015); and (4) patch-based (Habbecke and Kobbelt, 2006; Gallup et al., 2010). Methods based on depth maps, i.e., multi-view depth (MVD) map reconstruction, strive to reconstruct several depth maps image by image. Utilizing inputs that include multiple images and their internal and external camera parameters, the outputs are their respective depth maps, i.e., 2D array of 3D object points, where each pixel holds the distance from the principle plane to the corresponding object location (Hartley and Zisserman, 2003). Compared to other multi-view stereo methods, multi-view depth map reconstruction has been shown to be more adaptive to large-scale scenes due to its flexibility and scalability (Shen and Hu, 2014).

However, achieving accurate and complete depth map reconstruction for urban scenes continues to be a challenging task mainly because urban scenes often consist of regular shapes including flat objects, sharp corners, texture-poor regions, repetitive structures, and other complex features (Furukawa et al., 2009). Although the presence of these distinct geometric characteristics makes it possible to employ specialized constraints on final reconstruction, conventional MVD methods seem to be unequipped for such constraints. Furthermore, corners at walls and distinct roof or wall edges violate the common smoothness assumptions shared by most MVD methods. Many MVD methods (Gargallo and Sturm, 2005; Furukawa and Ponce, 2010) require depth discretization, which violates the planarity of plane structures; and the existence of noise and clutter in the results of common MVD methods often hampers their subsequent applications in urban scenes.

* Corresponding author at: Lyles School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA.
E-mail address: jshan@purdue.edu (J. Shan).

https://doi.org/10.1016/j.isprsjprs.2018.03.003
0924-2716/

This paper proposes a Planarity Constrained Multi-View Depth (PMVD) map reconstruction method for urban scenes. Using a planar geometry model as the local shape presentation, PMVD is able to reconstruct the depth and normal (vector) of the object simultaneously. The reconstruction starts with image segmentation and feature matching to initialize the local planes as optimization labels. Then, it utilizes an iterative optimization scheme, where each iteration is composed of candidate label selection, belief propagation, and outlier elimination. Based on the segmentation structure, PMVD extends the PatchMatch method (Barnes et al., 2009) to update the candidate labels at the beginning of each iteration. Compared to conventional methods, PMVD imposes genuine planarity constraints on depth map reconstruction in addition to simple smoothness. PMVD is evaluated qualitatively and quantitatively on several datasets of diverse scales and resolutions. The results demonstrate PMVD's ability to preserve the planarity of piecewise flat structures and the linearity of discontinuous edges while retaining good accuracy comparable to the state-of-the-art.

2. Related work

Many remarkable works about MVD methods have been published during the past few decades. Goesele et al. (2007) reconstructed a wide range of scenes from large, shared, multi-user photo collections available on the Internet by global and local view selection under a region growing process, and achieved high-quality outcome. Agarwal et al. (2011) achieved city-scale reconstruction with more than one hundred thousand images in less than a day. However, these methods did not consider structural regularities in urban scenes. As a result, structural priors were introduced where segmentation was utilized for detecting planarity structures (Holzmann et al., 2017). Pixels of the same segment were enforced to be on one plane (Sinha et al., 2009; Gallup et al., 2010). By assigning each segment an elevation, Duan and Lafarge (2016) achieved city-scale reconstruction from satellite images, but they could not handle oblique images. Due to the difficulty in selecting a proper scale for segmentation, these methods might overlook some details or create artifacts in urban scenes. Sinha et al. (2014) proposed a stereo algorithm which assigned each pixel to one of the local plane candidates generated by an iterative clustering step. However, the plane candidates in the aforementioned methods were determined before optimization and could not be changed afterwards.

There is an implicit assumption in the conventional MVD methods (Strecha et al., 2004; Goesele et al., 2007; Campbell et al., 2008) that the pixels in the support region of the center pixel have constant depth or disparity. These methods often use the fronto-parallel support window to compute image similarity, which is not suitable in urban scenes due to the prevalence of slanted surfaces. Recently, some novel strategies have been proposed. Bleyer et al. (2011) applied a slanted support plane with three parameters as the support region and obtained impressive sub-pixel results. Based on Bleyer et al. (2011), Shen (2013) and Bailer et al. (2012) proposed strategies that use one depth and two spherical coordinates to model the support plane. Zhu et al. (2015) adopted the support plane described by one depth and two depth offsets. A strategy proposed by Galliani et al. (2015) and followed by Schönberger et al. (2016) used a local plane in the Euclidean space as the support plane for the corresponding pixel.

Determination of the candidate depth set is an important problem for MVD methods. Conventional techniques often generate candidates with equal intervals in a certain depth range. The PatchMatch method (Barnes et al., 2009), a randomized algorithm for quickly finding approximate nearest-neighbor matches between image patches, has been widely used to generate candidate support planes from a continuous space (Zheng et al., 2014; Heise et al., 2015). There are two basic assumptions for PatchMatch: (1) some good patch matches can be found via random sampling; and (2) the natural coherence in the imagery allows propagation of such matches quickly to surrounding areas. By combining the PatchMatch method and max-product belief propagation (Felzenszwalb and Huttenlocher, 2006), PatchMatch Belief Propagation (PMBP) was applied in global stereo (Besse et al., 2013). Furthermore, an accelerated PMBP was proposed to handle critical computational bottlenecks, which achieved superpixel-based particle-sampling (Li et al., 2015). Galliani et al. (2015) achieved massively parallel multi-view extension of PatchMatch stereo on the graphics processing unit (GPU) for the propagation scheme. However, these successful efforts mainly utilized internet images, indoor images, and ground images and their formulation and applicability for aerial mapping purposes are unknown. Our proposed PMVD method extends the PatchMatch method to generate a candidate set of local planes with its segmentation structure during the depth map reconstruction for urban scenes.

MVD methods are image-based reconstruction techniques and therefore also have some fundamental image matching problems that need to be addressed. An image similarity metric has been used to measure the similarity of corresponding pixels in images with different perspectives and illumination. The squared differences or absolute differences of color values often were used in some real-time multi-view stereo (Hosni et al., 2013b; Galliani et al., 2015), while the normalized cross correlation and Census transform were adopted to consider the bias and gain changes across multiple images (Shen, 2013; Zheng et al., 2014; Li et al., 2015; Zhu et al., 2015). To achieve a reliable correspondence, aggregation of the image similarities was often necessary, i.e., applying a smooth filter over the image similarity space (Hosni et al., 2013a). The adaptive support-weight approaches improved the robustness of the similarity metric and achieved structure-preserving property (Hosni et al., 2013b; Hosni et al., 2011; Yoon and Kweon, 2006).

Occlusion is another problem for MVD methods. Determining the visibility of the object corresponding to a pixel in the reference image with respect to the target images (Zhu and Stamatopoulos, 2015) is necessary. Considering the noise distribution of the image similarity, Gargallo et al. (2005) estimated the visibility through a probabilistic model, which unfortunately was computation-demanding. Since the views whose scenes or objects are occluded return a low image similarity, the strategy for selecting the best views according to the image similarity metric has been applied widely (Kang et al., 2001; Galliani et al., 2015). Furukawa et al. (2010) handled this problem by restraining the image similarity and the angle between the ray and the surface normal vector.

We intend to address the problems or difficulties stated above, including avoiding the fronto-parallel effect, determining plane candidates in each optimization iteration, evaluating image similarity robust to occlusion, with the purpose to achieve planarity constrained multi-view depth map reconstruction. The detailed flowchart is illustrated in Fig. 1. The input is an image set $\mathcal{I}$ of N images as well as their camera parameters. Each image is in turn treated as a reference image denoted as I_r, and the rest of the images in $\mathcal{I}$ are denoted as target image set $\mathcal{I}' = \{I'_i\} = \mathcal{I} \setminus I_r$. We formulate the estimation of the depth map and the normal map as an optimization of the local planes for each pixel (Section 3.1). A sparse point cloud is first generated using feature matching and is assigned to the underlying homogenous segments of the reference image, which enables fitting a bundle of initial local planes as candidate labels. The local plane assigned to a pixel is then refined via an iterative optimization whose energy function consists of a similarity-based data term (Section 3.2) and a