ISPRS Journal of Photogrammetry and Remote Sensing 109 (2015) 47–61

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

# Accurate and occlusion-robust multi-view stereo

Zhaokun Zhu [a,b,*], Christos Stamatopoulos [c], Clive S. Fraser [c]

[a] College of Aerospace Science and Engineering, National University of Defense Technology, Changsha 410073, China
[b] Hunan Provincial Key Laboratory of Image Measurement and Vision Navigation, National University of Defense Technology, Changsha 410073, China
[c] CRCSI, Dept. of Infrastructure Engineering, University of Melbourne, Vic. 3010, Australia

ARTICLE INFO

*Article history:*
Received 23 January 2015
Received in revised form 25 August 2015
Accepted 27 August 2015

*Keywords:*
Multi-view stereo
Support window model
Visibility estimation
PatchMatch
Markov Random Field

ABSTRACT

This paper proposes an accurate multi-view stereo method for image-based 3D reconstruction that features robustness in the presence of occlusions. The new method offers improvements in dealing with two fundamental image matching problems. The first concerns the selection of the support window model, while the second centers upon accurate visibility estimation for each pixel. The support window model is based on an approximate 3D support plane described by a depth and two per-pixel depth offsets. For the visibility estimation, the multi-view constraint is initially relaxed by generating separate support plane maps for each support image using a modified PatchMatch algorithm. Then the most likely visible support image, which represents the minimum visibility of each pixel, is extracted via a discrete Markov Random Field model and it is further augmented by parameter clustering. Once the visibility is estimated, multi-view optimization taking into account all redundant observations is conducted to achieve optimal accuracy in the 3D surface generation for both depth and surface normal estimates. Finally, multi-view consistency is utilized to eliminate any remaining observational outliers. The proposed method is experimentally evaluated using well-known Middlebury datasets, and results obtained demonstrate that it is amongst the most accurate of the methods thus far reported via the Middlebury MVS website. Moreover, the new method exhibits a high completeness rate.


## 1. Introduction

Image-based modeling and reconstruction is the process of extracting accurate three-dimensional (3D) information of unknown objects or scenes from imagery. It is essential for many photogrammetric measurement applications, such as product quality inspection, vision metrology, cultural heritage documentation, traffic accident reconstruction and mobile mapping. 3D reconstruction from stereo and multi-view imagery has been given impetus over recent years due in large part to the decreasing costs and higher resolution of consumer-grade digital cameras. A more important factor in the growth of applications of image-based modeling, however, has been the new developments in image matching algorithms that have allowed faster, more robust and more accurate reconstruction.

3D reconstruction can be grouped in two major categories namely, sparse and dense reconstruction. Sparse reconstruction is usually performed via detector-based matching of common scale invariant features (Bay et al., 2008) between images. This facilitates the determination of both image orientation and a sparse reconstruction, desirably via established photogrammetric methodologies. It is noteworthy that 'sparse' may nevertheless imply 1000s of extracted 3D object points. This technique is commonly referred to as Structure-from-Motion (SfM) (Crandall et al., 2013) and it has been successfully applied to large sets of both organized and unorganized images. On the other hand, dense reconstruction methods aim to create a complete model of a scene by extracting a 3D coordinate for every pixel of an image, thus leading to object point clouds of millions of points. A prerequisite for practical dense reconstruction is the provision of images with known interior and exterior orientation and thus it is not uncommon nowadays to have sparse and dense reconstruction performed as sequential operations.

Dense reconstruction methods can be further divided into binocular stereo and multi-view stereo (MVS) methods. Binocular stereo (Scharstein and Szeliski, 2002) aims to produce a dense disparity map from a pair of images, whereas MVS methods aim to utilize multitudes of images. According to the taxonomy of Seitz

* Corresponding author at: College of Aerospace Science and Engineering, National University of Defense Technology, Changsha 410073, China. Tel.: +86 186 7087 7752.

http://dx.doi.org/10.1016/j.isprsjprs.2015.08.008
0924-2716/

et al. (2006), MVS methods can be further categorized in terms of how the object is represented, for example via voxels, level-sets, polygon meshes or depth maps. MVS methods in which the object is represented by depth maps are also referred to as multi-view depth map estimation (MVDE) methods by Zheng et al. (2013).

One fundamental image matching problem that both local binocular stereo and MVDE methods have to deal with is how to optimally select the support window along the epipolar line of the support (matching) image to ensure photo-consistency, given a reference window in the reference image. The distinction for MVDE methods is that support windows have to be selected for all possible support images. For the sake of simplicity, many methods use the same window size as the reference window, assuming that the reconstructed surface is locally fronto-parallel. Such an assumption is easily violated in many situations, so various local binocular stereo methods have proposed alternative methodologies. These include the methods of adaptive-window (Veksler, 2003), multiple-window (Bobick and Intille, 1999), segmentation-based window (Wang et al., 2004), adaptive support-weight (Hosni et al., 2013) and Disparity Space Image (DSI) plane induced support window (Bleyer et al., 2011). Many MVDE methods, e.g., that proposed by Goesele et al. (2006), use the same size reference window, but further problems can arise with this approach since many pixels of the support window do not necessarily correspond to the same object point at all, resulting in a deviation of the computed photo-consistency measure. Thus, more sophisticated support window calculation is also warranted for MVDE. However, the complexity increases because instead of only one depth parameter along the line of sight, there are now multiple parameters involved in the calculation of the depth estimate.

Another challenge for MVDE methods is the determination of the visibility of each pixel of the reference image with respect to the support images. More specifically, it is the process of finding which support images satisfy the photo-consistency measure for every pixel of the reference image so that this information is taken into account during the depth estimation process. This step is quite important, as the introduction of erroneous matches will lead to incorrect depth calculation and consequently outliers. Conversely, use of only a small number of estimations in order to minimize the outliers, e.g., the nearest support images, is also not desirable as redundant observations are essential in providing better geometry and consequently increasing the accuracy and reliability of the depth estimation. Although outliers can be detected and filtered by enforcing depth-consistency with other depth maps, the final completeness of the reconstruction can be adversely affected by the presence of outliers. Thus, the visibility determination is crucial to achieving both high completeness and accuracy in the 3D reconstruction.

The problems of support window selection, visibility determination and outlier detection in image-based 3D reconstruction are addressed in this paper, where a novel MVDE-based method for the estimation of both depth and surface normal for each pixel is proposed. The reported approach relies on an accurate visibility estimation method that is initially calculated with the purpose of efficient outlier filtering. Thus, high accuracy depth and surface normal estimates are anticipated as all possible redundant observations can be correctly recognized and utilized. In addition, so as to allow for faster computation times, the proposed methodology is developed in such way that parallel processing of each image is possible. The proposed MVDE approach is evaluated using well-known Middlebury MVS benchmark datasets, and results obtained show the method to be amongst the most accurate reported in accommodating occlusion-challenging scenes. It is also capable of producing a 3D reconstruction with a very high completeness rate.

## 2. Related work

Selected MVDE approaches, of which there are many in the computer vision literature, will first be discussed to provide a background into the MVDE-type approach adopted here. A classic MVDE approach is that reported by Goesele et al. (2006), in which normalized cross-correlation (NCC) is used as the photo-consistency measure. The support window model is simply a fixed-size square window without any surface normal information for the associated 3D object points, while the visibility of each pixel is determined based on both the computed NCC and a comparison against a fixed threshold. Such an approach can often lead to erroneous results as in some cases a pixel can have a high NCC value even if it is occluded.

In aiming to handle large online 'community' photo collections, Goesele et al. (2007) further proposed a region growing MVDE method that takes reconstructed sparse feature points from SfM as seeds and iteratively grows surfaces from them. Additionally, for the first time, a support window of non-fixed size was introduced. This window is derived by a 3D support plane modeled in the image domain by a depth and two depth offsets. Two photo-consistency measures are used at the same time, namely the sum of squared differences (SSD) and NCC. The SSD is employed only for parameter optimization, while NCC is used for calculating both confidence and convergence, as well as determining visibility in a similar way to that reported in Goesele et al. (2006). Optimization of both the depth and the two depth offsets is solved via a multi-photo geometrically constrained least-squares matching (MPGC) method (Baltsavias, 1991), one of the drawbacks of which is that it requires good initial parameter estimates in order to achieve fast convergence. Otherwise MPGC may either achieve convergence slowly, oscillate or even diverge (Gruen, 1996). To deal with these issues, a specific yet complicated mechanism was proposed, yet the completeness and accuracy of the reconstruction remains heavily dependent upon the SfM seed points.

In the method of Strecha et al. (2006), the depth and visibility of each pixel of the reference image are jointly modeled as latent variables in a hidden Markov Random Field (MRF) where smoothness is incorporated on neighboring pixels for both depth and visibility. Their inference, along with the stochastic model parameters, is calculated by an Expectation-Maximization (EM) algorithm. Unlike other MVDE methods, the support window is not necessary as the algorithm is able to consider only pixel-wise color discrepancies due to the incorporation of smoothness terms. However, it is assumed both that every matching pixel has the same color, without error, and that the color discrepancy should follow a Gaussian distribution with zero mean, something that is usually violated in practical situations. Further drawbacks of this method include high memory-consumption due to the need to store the visibility configuration for each pixel, with memory requirements increasing exponentially as the number of images increases.

The use of several locally optimal depth hypotheses for each pixel has been proposed by both Campbell et al. (2008) and Hu and Mordohai (2012). Campbell et al. (2008) select final depth estimates based on a discrete MRF model without the presence of any other depth maps, whereas Hu and Mordohai (2012) adopt a depth estimation based on depth consistency with other depth maps. While the former of these two approaches uses a fixed-size square support window, as with the method of Goesele et al. (2006), the latter uses a plane-sweeping stereo method similar to that reported by Gallup et al. (2007) where four different support windows with corresponding 3D object surface normals are considered. Both methods involve finding several local extrema of the photo-consistency measure, which requires sampling the entire