Continuous Depth Estimation for Multi-view Stereo

Yebin Liu Xun Cao Qionghai Dai Wenli Xu
Automation Department, Tsinghua University, Beijing 100084, China
{liuyb02,cao-x06}@mails.tsinghua.edu.cn,{qhdai,xuwl}@tsinghua.edu.cn

Abstract

Depth-map merging approaches have become more and more popular in multi-view stereo (MVS) because of their flexibility and superior performance. The quality of depth map used for merging is vital for accurate 3D reconstruction. While traditional depth map estimation has been performed in a discrete manner, we suggest the use of a continuous counterpart. In this paper, we first integrate silhouette information and epipolar constraint into the variational method for continuous depth map estimation. Then, several depth candidates are generated based on a multiple starting scales (MSS) framework. From these candidates, refined depth maps for each view are synthesized according to path-based NCC (normalized cross correlation) metric. Finally, the multiview depth maps are merged to produce 3D models. Our algorithm excels at detail capture and produces one of the most accurate results among the current algorithms for sparse MVS datasets according to the Middlebury benchmark. Additionally, our approach shows its outstanding robustness and accuracy in free-viewpoint video scenario.

1. Introduction

Multi-view stereo (MVS) aims to reconstruct watertight 3D model from multiple calibrated photographs of a realistic object. Because of many potential applications, i.e. industrial design, characters modeling for films and electronic games, demonstration of cultural relics, and commercial advertisement, MVS has drawn more and more attentions in recent years. Although many algorithms have been developed for this problem, efforts still have to be made to achieve a 3D modeling with both high efficiency and high quality.

Following the taxonomy of Seitz et al. [22], MVS algorithms can be generally classified into four categories: 3D volumetric approaches [15, 28, 26, 27], surface evolution techniques [21, 8, 31], feature extraction and expansion algorithms [29, 9, 11], and depth map based methods [10, 16, 5, 30]. Among these four classes, depth map based approaches make up an important part of those top performers on standard evaluation tests [2]. Generally, such methods involve two separate stages. First, a depth map is computed for each viewpoint using binocular stereo. Second, the depth maps are merged to produce a 3D model. This two-stage strategy offers great flexibility to embed different techniques into the whole reconstruction pipeline. In these methods, the estimation of the depth maps is crucial to the quality of the final reconstructed 3D model. Compared with traditional binocular stereo problems [7], an accurate pixel-level processing is required in MVS to recover a continuous and precise model surface. In this situation, discretized global optimization algorithms such as graph cuts and belief propagation, which are widely adopted in binocular stereo problems, will lead to quantization errors during the discretion of depth values as well as huge memory requirement.

To guarantee pixel-level optimization, Goesele [10] uses window based matching technique to compute color consistency between neighboring views along the epipolar line. Because only the pixels with high color consistency are adopted, large part of the object surface will not be reconstructed when cameras are sparse. Recent depth map based MVS approaches [19, 16, 5] are also based on window matching schemes to assign one or multiple depth candidates for each pixel. By introducing the outlier removing techniques, promising depth maps can be obtained without large memory requirement. However, because of the distorted matching windows caused by slanted projection of surface patches, these methods and some of the feature growing approaches [29, 17, 11] either endeavor to rectify distorted matching windows or search the possible slanted angles for accuracy matching. Moreover, it is time-consuming for these methods to achieve sub-pixel matching precision for a high accurate surface recovery.

In this paper, we propose a variational depth map estimation technique aiming at high quality MVS reconstruction. Continuous variational approach has the ability to overcome the above mentioned limitations such as discretization errors, memory consumption and distorted matching windows. Such continuous formulation naturally enables rotation invariant photo-consistency matching. For MVS

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

problems, we find that these properties are especially suitable for capturing continuous surface details. In addition, a multiple starting scales (MSS) technique is proposed to generate multiple depth map candidates from different starting scales. The presence of multiple candidates can prevent some occasional over-smoothness and avoid troublesome local minima in single depth map. Then refined depth map is synthesized, i.e., multiple depth maps are fused through a selection and cleaning process. Thanks to the continuity of the candidate depth maps, the data term can be accurately measured using patch-based NCC metric [29].

Figure 1 illustrates the whole pipeline of our reconstruction algorithm. Compared with traditional depth map based approaches, the most obvious unique of our proposed algorithm lies in the variational depth map estimation and MSS depth map synthesis technique.

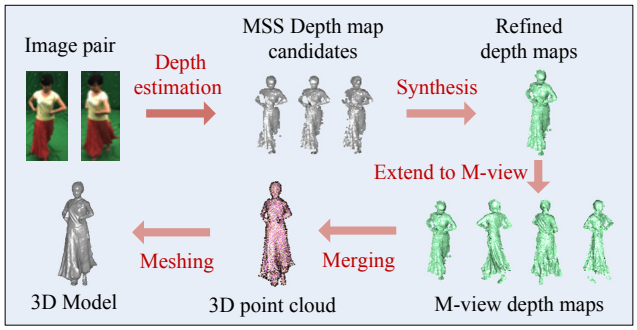

Figure 1. Reconstruction pipeline of our proposed method.

In particular, our MVS pipeline has the following advantages:

• Combined with visual hull and epipolar constraint, variational technique can provide continuous surface with delicate details without the complex pixel by pixel window matching.

• The fusion of depth maps on each individual view is based on a MSS technique. Both the high frequency and low frequency errors are damped through the selection of multiple starting scale iteration results.

• Continuous depth-map candidate allows accurate NCC measurement on each pixel to choose the optimal depth. Such accurate matching circumvents distorted matching windows due to the slanted surfaces.

In the following, we first review related work in Section 2. We present the variational depth map estimation algorithm and its MSS framework in Section 3. In Section 4, we describe the whole merging procedures including the fusion of continuous depth map candidates for individual camera view and the merging of multi-view depth maps to produce 3D point cloud. Finally, we show the experimental results in Section 5 and conclude the paper in Section 6.

2. Related Work

The first free viewpoint video studio, Visualized Reality^{TM} [25], adopts a depth map merging reconstruction scheme. Model reconstruction did not perform well at that time because of ambiguous stereo matching and rough surface fusion. After this work, lots of 3D model reconstruction algorithms turned to use the discrete global optimization techniques [12, 28, 13] to produce 3D surface models. In recent years, Goesele [10] revisits depth map based approaches using pixel by pixel window matching technique to retrieve the high fidelity matching pixels and merge the resulted depth maps using volumetric meshing algorithms. Bradley [5] improves this works by increasing the number of matching pixel pairs using scaled window matching technique and depth map filtering technique. Campbell [19] enhances the quality of depth-maps by extracting multiple depth candidates for each image pixel, and then impose global optimization algorithm to simultaneously remove outliers and achieve depth map smoothness. These methods are all pixel-level matching technique for discrete depth map generation. As for works that concentrate on multi-view depth maps merging, Merrell [16] addresses the problem of real-time depth map merging via GPU technique. Zach [30] computes the depth maps to create a 3D model using total variation regulation and L1 norm to measure data drift.

Among all the techniques used in MVS, photo consistency measure is crucial to the 3D reconstruction performance. Rectangular window matching in image space [28, 27, 10] is not accurate due to the camera projection transformations of the object surface. Patch based matching [29] in scene space is much more reasonable by adding the surface normal and position information. However, it is rather time-consuming to traverse all normal and position situation for each surface patch. Furukawa [29] selectively optimize the normal and position of patches corresponding to salient image features to save the computation time while maintaining surface accuracy and completeness by patch propagation algorithm. This algorithm achieves extremely high performance on lots of MVS datasets. Scaled window matching technique proposed by Bradley [5], and plane fitting optimization suggested by Habbecke [11] are also schemes that can rectify matching window or distorted color consistency matching in acceptable time. Our work proposes a novel matching solution based on continuous depth map candidates.

In the area of optical flow estimation, variational approach has demonstrated its strength by presenting good accuracy in developing continuous correspondence between motion images [4]. The filling-in effect creates dense correspondennce maps with sub-pixel level precision by propagating information over the entire image domain. Based on these advantages, Slesareva [20] tries to add epipolar con-

straint to the traditional optical flow scheme for binocular stereo problems. In his later work [23], this scheme has been extended to multiple ortho-parallel views aiming for more accurate depth maps. Variational approach has also been successfully applied in binocular scene flow scenario, in which depth and motion are jointly optimized. In MVS area, Kolev [14] proposed a continuous global optimization algorithm parallel to discrete volumetric graph cuts [28]. Although it is based on variational technique, its solution is in the 3D space, while ours is based on the 2D image space.

3. Continuous Depth Estimation

As mentioned above, the depth estimation is important to the final performance of MVS system. In this section, we first describe the variational approach applied for continuous depth map generation from binocular image pairs. Then we introduce the MSS technique adopted to generate the multiple depth map candidates for the final depth map synthesis. This strategy is inspired by the observation that the optima of low frequency image components are easier to find when we rescale the image to a coarse level. In addition, by offering more depth candidates, this MSS method can achieve a more robust and accurate depth result.

3.1. Variational depth estimation

Variational approach has made tremendous progress on optical flow problems[4], which is very similar to stereo matching. Both of these two problems are dedicated to find the corresponding pixels among picture frames. Variational approach shows potential for 3D model reconstruction. First, watertight objects (e.g. humans and toys) have surfaces that tend to be continuous and free from abrupt changes. Second, silhouette information of the object provides a reduced searching space and a favorable initial value for variational optimization. Moreover, because intrinsic parameters and camera poses are known, epipolar line can reduce the searching space from two dimensions to only one dimension. This simplification is illustrated in Figure 2.

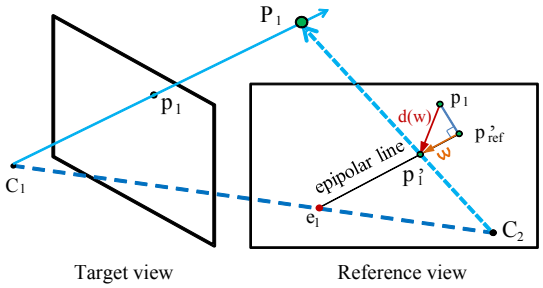

Figure 2. Restriction on displacement from 2D to 1D.

Figure 2 shows that according to the epipolar geometry, we can search for the corresponding point just on the epipolar line, instead of two dimensional searching in optical situation. Originally, if we want to find the corresponding point of p_1 in the neighboring view, for example, p_1' in the reference view, we must use displacement in both x and y directions. At present, we can just search for the corresponding point on the epipolar line. We first get the so-called "anchor point" p'_{ref} by drawing a line trough p_1 which is perpendicular to the epipolar line. Then, the searching displacement is defined based on w, which is the displacement from p'_{ref} to p'_1. Since the distance between p_1 and p'_{ref} is fixed for each pixel, the original 2D displacement in optical flow problem can be replaced by $d(w)$. With the introduction of $d(w)$, the depth maps can be computed by minimizing the target energy functional which includes both the local feature matching and the global smoothness terms:

$$E(w) = E_D(w) + \alpha E_S(w), \tag{1}$$

where

$$\begin{aligned} E_D(w) = \int_\Omega \psi_D(|I_r(p + d(w)) - I_t(p)|^2 \\ + \gamma|\nabla I_r(p + d(w)) - \nabla I_t(p)|^2)dxdy, \end{aligned} \tag{2}$$

and

$$E_S(w) = \int_\Omega \psi_S(|\nabla w|^2)dxdy. \tag{3}$$

Here, I_r is the reference image and I_t is the target image. The data term contains two parts: the first part assumes the color value consistency in each view and the second part models the constancy of the spatial image gradient. The introduction of gradient term ∇I makes the approach more robust to varying illumination. In the meantime, the spatial gradient can preserve the edges well. Since the data constraint is not always accurate, e.g. because of occlusions, brightness changes or noises, we apply the robust function ψ proposed in [6]. The smoothness term is computed by taking the interaction between neighboring pixels into consideration. In this case, it is designed to penalize the total variation of the flow field.

The problem of finding the functions w that minimize the target variational energy function can be converted to the minimization of Euler-Lagrange equation. For simplicity, we use the following abbreviations for derivatives and differences:

$$\begin{aligned} I_x := \partial_x I_r(p + d(w)), I_{xy} := \partial_{xy} I_r(p + d(w)), \\ I_y := \partial_y I_r(p + d(w)), I_{yy} := \partial_{yy} I_r(p + d(w)), \\ I_z := I_r(p + d(w)) - I_t(p), I_{xz} := \partial_x I_r(p + d(w)) - \partial_x I_t(p), \\ I_{xx} := \partial_{xx} I_r(p + d(w)), I_{yz} := \partial_y I_r(p + d(w)) - \partial_y I_t(p). \end{aligned} \tag{4}$$

Therefore, the Euler-Lagrange equation is expressed as:

$$\begin{aligned} &\psi_D^{'}(I_z^2 + \gamma(I_{xz}^2 + I_{yz}^2)) \\ &\times((I_z I_x i + \gamma I_{xz} I_{xx} i + \gamma I_{yz} I_{xy} i) + (I_z I_y j + \gamma I_{xz} I_{xy} j + \gamma I_{yz} I_{yy} j)) \\ &-\alpha div(\psi_S^{'}(|\nabla w|^2)\nabla w) = 0. \end{aligned} \tag{5}$$

Here, $(i, j)^T$ is the unit normal vector of direction of epipolar line corresponding to pixel p. The above equation is nonlinear in both data term and smoothness term, and can be solved by a multi-resolution strategy with two nested fixed point iterations to remove the nonlinearities in the equations. A coarse-to-fine strategy with down-sampling factor η is used. Such solving strategy can be theoretically justified as an approximation to the continuous energy functional. The initial values w_0 of w for each pixel in the target image should be pre-computed based on the visual hull. Moreover, the directions of epipolar lines and the anchor points for all pixels should also be pre-computed and saved as a direction map and anchor map respectively. During multi-scale iteration, these maps serve as look-up tables for the down-sampled pixels to check their epipolar directions and anchor positions.

In summary, variational technique tends to get stuck into local optima. However, when it is combined with visual hull and epipolar constraint, which provide a satisfactory initial value and a restricted feasible space for optimization, the result will have a big chance to achieve a high quality depth map, although it may not be the global optima. Compared with traditional window based matching technique, variational formulation naturally enables rotation invariant matching, and such invariance permits matching without consideration of surface perspective distortions. Therefore, details can be perfectly recovered on distinctive regions. However, variational formulation cares only intensity and gradient consistency for neighbor pixels while neglecting the region segment based matching characteristic, which will degrade the matching accuracy.

3.2. Multiple starting scales (MSS) framework

Due to the inherent difficulties such as noise, occlusion errors, lack of textures, texture repeating and some other uncertainties, errors may occur during the variational matching computation. Thus, a single coarse-to-fine iteration is not enough. To improve the accuracy and robustness, we propose a MSS (Multiple starting scales) framework which starts the variational depth estimation at various scale levels. This method offers more diverse depth candidates than a single coarse-to-fine iteration.

We use *templeSparseRing* and *dinoSparseRing* to illustrate the advantage of MSS. The resolution of these input images is 640 by 480. Figure 3 shows the depth map reconstructed under different starting scales. From the left to the right of each row, the starting iteration level is getting coarser. For example, in the temple sequence (a)∼(d), we rescale the image using factor η ($0< \eta <1$) by 5, 10, 15, 20 times respectively as a starting level. Namely, the result of Figure 3(a) is obtained through a coarse-to-fine iteration from coarse resolution ($640\cdot\eta^5$, $480\cdot\eta^5$) to fine resolution (640, 480). We mark the regions which are well reconstructed and not-well reconstructed in Figure 3.

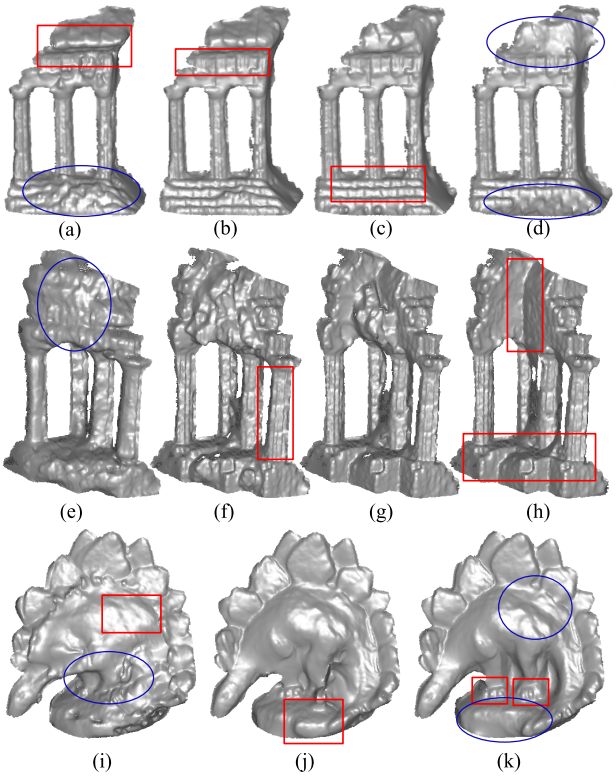

Figure 3. Influence on reconstructed results by different starting resolution level. All the three examples are fine-to-coarse from the left to the right. The red rectangles mark the regions which have been well reconstructed, while the blue circles sign the regions which fail to recovered. (a)∼(d):Temple view 1; (e)∼(h):Temple view 2; (i)∼(k):dinosaur example.

When the depth computation is iterated from fine level, the surfaces that are approximate to the initial value (visual hull) will be recovered (see the roof and the pillars of temple in (a) and (f)) while some large displacements like deep concaves may not be carved. See the top of the temple in (e) and the foot of dinosaur in (i). The failure on concaves verifies the fact that,without iteration from coarse level, it is difficult to find the long distance correspondences and results can be easily stuck into local minima.

Conversely, when iterating from coarse level, the salient shape information like the deep concaves can be perfectly reconstructed (see the concave wall in (h) and the foot of dinosaur in (k)). This success is attributed to the continuous property of variational technique as described above. However, for textureless regions, the results are prone to be over-smooth or erroneous. (see the roof of the temple in (d) and the collapsed round table in (k). The reason for this unpleasant reconstruction lies in the fact that the details degraded due to the down-sampling strategy.

At last, for some regions with periodic texture such as the steps in (a)∼(d), different starting levels present various

reconstruction results.

From these analysis, it is incomplete to fix on a particular starting level, while the combination of depth map candidates at different starting scales has the potential to achieve comprehensively satisfactory reconstruction.

4. Merging and Meshing

Our merging can be divided into two stages: depth map synthesis and multi-view merging. Depth map synthesis is to generate a high quality depth map for each camera view based on the MSS depth map candidates, while multi-view merging aims at producing an as-clean-as-possible point cloud using the multi-view refined depth maps.

4.1. Depth map synthesis

We use the patch based color consistency metric to retrieve a depth value from the multiple candidates for each pixel. Thanks to the continuous property provided by the variational depth map, accurate patch based matching for pixels is possible, as illustrated in figure 4.

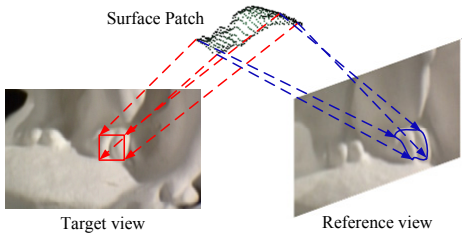

Figure 4. Patch-based NCC measurement based on continuous variational flow and surface patch.

For the measure of pixel photo consistency on a specific depth map candidate, corresponding 3D point with its neighbor points can be traced and then clustered to be a surface patch. Photo consistency metric such as normalized cross correlation (NCC) on the projection regions can then be ideally computed based on this surface patch. To circumvent point cluster operation in 3D, NCC matching can be conveniently defined in image space as:

$$NCC(p_i, s) = \frac{\sum_{j=1}^{N^2}(n_j - \overline{n}) \cdot (f_s(n_j) - \overline{f_s(n)})}{\sqrt{\sum_{j=1}^{N^2}(n_j - \overline{n})^2} \cdot \sqrt{\sum_{j=1}^{N^2}(f_s(n_j) - \overline{f_s(n)})^2}}. \tag{6}$$

Here p_i is the target pixel in the primary view, s is the index of candidate depth maps, and n_j is local neighborhoods of size $N \times N$ in the target image. $f_s(n_j)$ is the pixel region corresponding to n_j in the reference images obtained by the variational flow f_s. $\overline{n}$ and $\overline{f_s(n)}$ represent the intensity averages over the two regions.

Suppose that the number of depth map candidates is S. Generally, s^* that satisfies

$$s^* = \arg \max_{s=1,2...S} NCC(p_i, s) \tag{7}$$

will correspond to the accurate depth. To improve accuracy and reduce the number of outliers in the point cloud corresponding to the final depth map, we remove points that may be erroneous as follow. If a point whose neighbor number is small, it may be a disconnected point and should be filtered out. We use a fixed radius to compute the local neighborhoods for the point cloud. Points with neighbor number lower than $0.5n$ should be removed. Here, n is the average neighbor number of all the points. Moreover, we add constraint to remove the points whose angles between their normals and the view-vector (the vector from the point to the view camera) are larger than 45 degree. This is because the larger angles means these points may not be accurate and their counterparts in other cameras are more favorable. Figure 5(a)illustrates two views of the depth map synthesis results for the *dinoSparseRing* data set.

In summary, our depth map synthesis combines both the advantages of the patch based NCC calculation in PMVS [29] and the multiple hypothesis optimization (MHO) [19]. PMVS traverses all feasible patch normal and patch positions for each feature pixel, while continuous depth map provides natural patches and involves only the choosing of the best from all available candidates. On the other hand, compared with MHO, candidates in our method are patches, which is more accurate for stereo matching than discrete 3D points used in MHO.

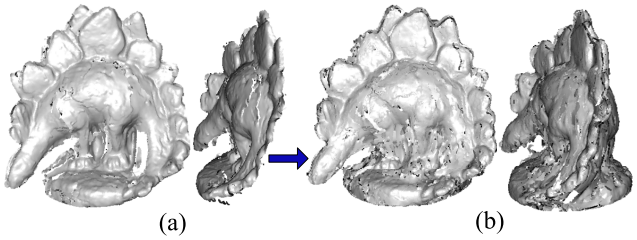

Figure 5. Depth map synthesis result (a) and multi-view depth map merging result (b) on *dinoSparseRing*. Results are obtained by shading the point clouds corresponding to the depth maps.

4.2. Multi-view merging and point cloud meshing

All multi-view depth maps are combined together to get a huge point cloud with each point containing information including position, normal direction, detecting camera and color consistency value. Attributing to the high qualities of our multi-view depth maps, the merging process can be time-efficient in our work. We remove outliers by detecting the conflicting point pairs. The following criteria define two points as conflicting point pair: 1) captured from different cameras, 2) project to the same position on a particular camera view, but their distance is smaller than threshold, 3) to the projected camera view, they have the same sign of normal direction. Our removing algorithm assumes that for any pair of points satisfying with the above criteria, the

point with lower color consistency value in the conflicting pair should be considered a noise point and will be removed. Here, the projected camera view is allowed to be a virtual camera view to increase the conflicting point detecting rate. Figure 5(b) shows two views of the merged point cloud.

Finally, the merged point cloud is meshed based on the Poisson surface reconstruction [18] because of its ability to generate a watertight mesh and its robustness to both noise and non-uniform sampling rate. To rectify reconstructed vertices that lie outside the visual hull, we project each of such vertices back to the visual hull according to their inwards normal directions.

5. Results

5.1. Implementation

The variational depth estimation module is implemented based on Brox's optic flow method [6]. Table 1 lists the parameters used in our algorithm. For the two Middlebury datasets, since the true surfaces are far from their corresponding visual hull, we output 4 kinds of starting scales (s = 5, 10, 15, 20) for depth map synthesis. As for our captured datasets, only 3 starting scales (s = 1, 4, 8) are necessary.

Table 1. Summary of the parameters used in the experiments.

Parameters	Value
Reduction factor η in variational flow	0.9
Outer fixed point iter. in variational flow	10
Inner fixed point iter. in variational flow	2
SOR iter. in variational flow	5
$\sigma/\alpha/\gamma$ in variational flow	1.8/80/100
Stereo window matching size	7×7

The main computation time is spent on the variational depth estimation. For low resolution images (e.g. Milddlebury datasets with a resolution of 640×480), this process is fast. Our algorithm is further accelerated because foreground extraction makes large areas of background unnecessary in the variational iterations. Besides, since pixel correspondences are known before NCC calculation, the complexity of the depth synthesis is also very low. Finally, the point cloud merging module is time efficient. All these facts make our algorithm feasible in a satisfactory time.

5.2. Reconstruction results and evaluation

The quantitative results of our algorithm are shown in Table 2. At the moment these results are submitted, the accuracy and completeness measurements of these two datasets both rank top 4. Figure 6 shows our final reconstruction results of these two datasets. Details such as the foot of the dinosaur and the pillars of the temple have been ideally reconstructed.

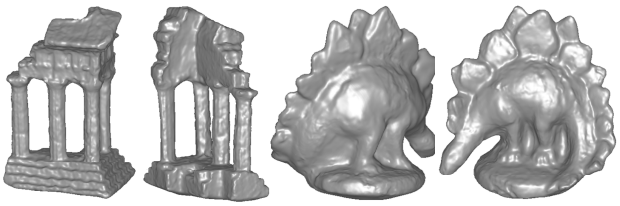

Figure 6. Reconstruction results of *dinoSparseRing* and *templeSparseRing*.

Table 2. Results for Middlebury datasets. Accuracy is measured in millimeters, completeness as a percentage of the ground true model, time in minutes. Both the accuracy and completeness of these two datasets rank top 4 of the Middlebury evaluation.

DinoSparseRing			TempleSparseRing		
Acc.	Comp.	Time.	Acc.	Comp.	Time.
0.51	98.7	28	0.65	96.9	23

Our algorithm is also effective and efficient for free-viewpoint video (FVV) datasets captured by multi-camera array. Compared with the static multi-view video datasets, FVV datasets suffers from the problems such as low resolution, unfavorable color synchronization, image noises and invisible surface regions. We have setup a multi-camera 3D studio to capture multi-view videos for human actors. These multi-view datasets consist of 20 views evenly spaced on a ring. The spatial resolution of each image is 1024×768. Our datasets are pretty challenging because of the presence of the above mentioned problems to some extent. These datasets are available on [1]. Figure 7 illustrates our reconstruction results on different kinds of dressings and poses. The red rectangles on the models mark the challenging regions that we have successfully reconstructed. For example, the hair on the shoulder is still visible in (a), and the wrinkle of the back can be seen clearly in (d). Such recovered details are difficult to handle for graph cuts methods [28] because of their limited grid resolution and the discrete approximation of the smoothness term.

Figure 8 additionally emphasizes a comparison with volumetric graph cuts (Figure 8(a)), patch based MVS [29] ([29], Figure 8(b)) and our proposed methods (the first row of Figure 9) when applied on our captured datasets. The volumetric graph cuts is implemented based on [27], and we use the PMVS software [3] to obtain the patch based MVS results. Here, the reconstruction results by graph cuts look slightly over-smooth. For example, the space between the actor's legs in the first two images can not be carved out. These artifacts are caused by the discrete approximation of the smoothness term and the inaccurate square window matching. PMVS is perfect for the static high quality MVS datasets, but it relies on the accurate reconstruction of the feature patches to propagate the surface. For area without enough texture information, such as the black hair and the thin arms, PMVS cannot detect enough confident fea-

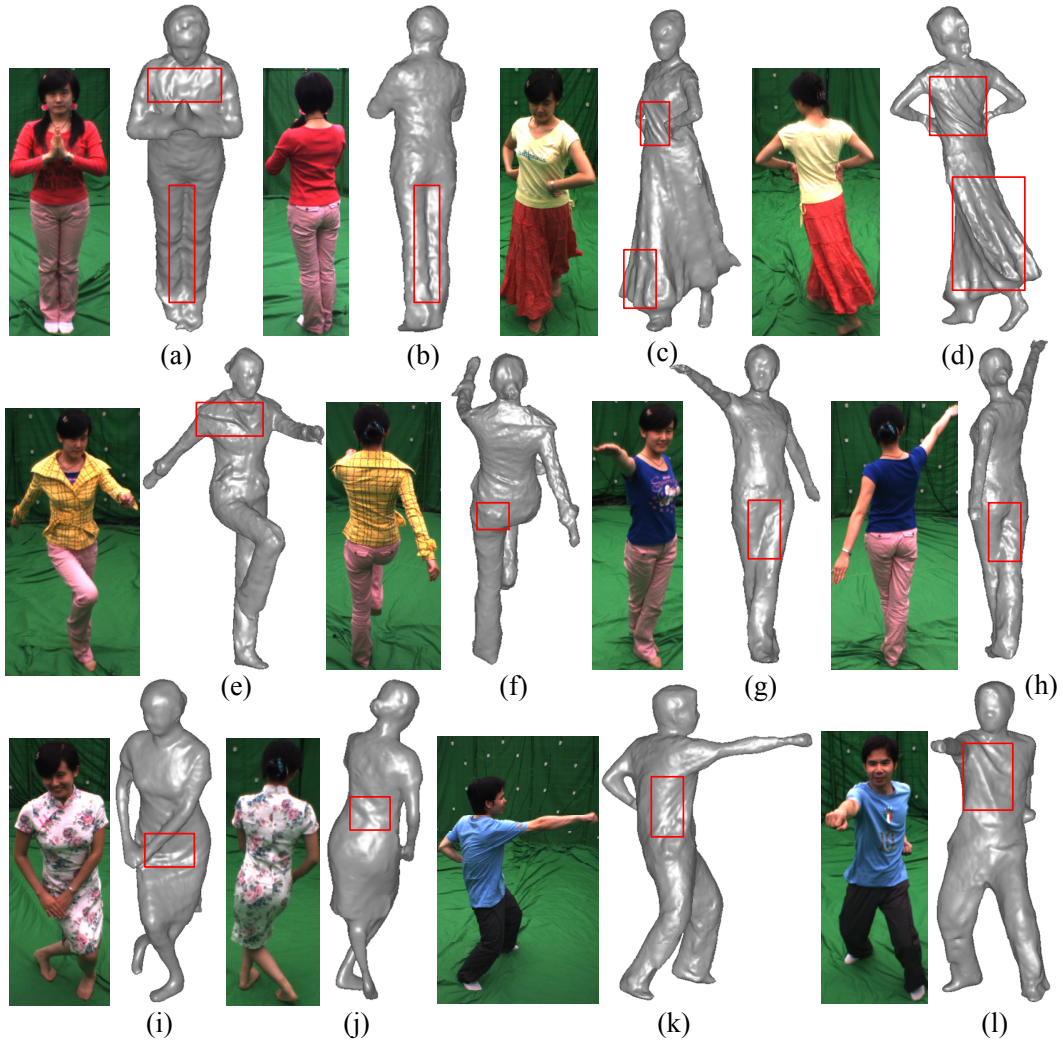

Figure 7. Reconstruction of free-viewpoint video datasets. The grey images are reconstructed models and the color ones are input images. Red rectangles mark the challenging regions that we have successfully handled.

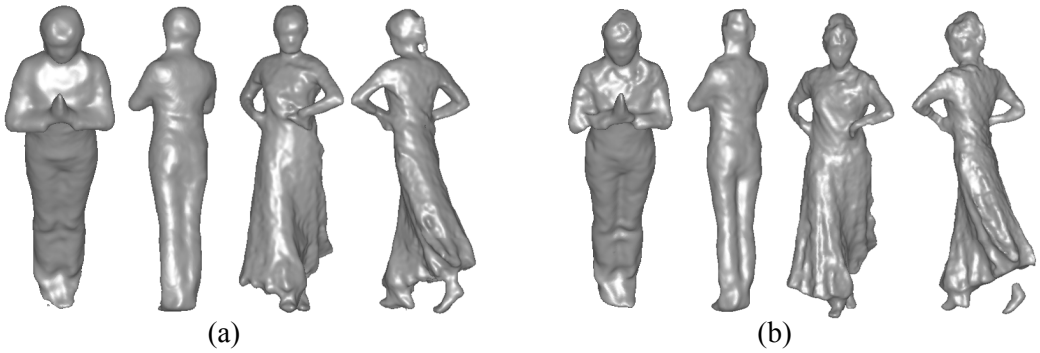

Figure 8. Reconstruction result using: (a) Graph cuts method similar to [27], (b) patch based MVS (PMVS) [29].

tures, thus the results may not be robust. In contrast, our continuous depth map based method can achieve accurate capture of features, nice reconstruction on smooth regions and robustness on challenging datasets.

6. Conclusion

In this paper, a novel multi-view stereo algorithm using an estimation-synthesis-merging framework on continuous

depth maps is proposed. By introducing variational flow technique to MVS area and combining it with visual hull information and epipolar constraint, we are able to capture 3D surface details conveniently, without the need of pixel by pixel window matching procedure. In addition, continuous depth map enables accurate photo-consistency measurement, and such advantage further promotes the depth map synthesis strategy to choose the optimal depth value from the multiple depth candidates. It is the combination of these advantages that allows robustness and high quality reconstruction on both static multi-view and motion multi-camera datasets. The multi-view depth map merging module in our algorithm is plain, which leaves space for future performance improvement.

Acknowledgments: This work was supported in part by the Distinguished Young Scholars of NSFC No.60525111, the 863 program, No. 2007AA01Z332 and the Intel-Tsinghua Cooperation Fund. We thank Bennett Wilburn in Microsoft Research Asia for the suggestion on multi-camera system construction, S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski for the temple and dino datasets and evaluations.

References

[1] http://media.au.tsinghua.edu.cn/cmvs.jsp. 6

[2] http://vision.middlebury.edu/mview/. 1

[3] http://www.cs.washington.edu/homes/furukawa/. 6

[4] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *Proceedings of the IEEE International Conference on Computer Vision, ICCV*, 2007. 2, 3

[5] D. Bradley, T. Boubekeur, and T. Berlin. Accurate multi-view reconstruction using robust binocular stereo and surface meshing. *CVPR*, 2008. 1, 2

[6] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optic ow estimation based on a theory for warping. *ECCV*, 3024:25–36, 2004. 3, 6

[7] D.Scharstein and R.Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Intl. Journal of Computer Vision,IJCV*, 47(1). 1

[8] C. H. Esteban and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, 2004. 1

[9] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. *ICCV*, 2007. 1

[10] S. M. Goesele, B. Curless, and S. M. Seitz. Multi-view stereo revisited. *CVPR*, pages 2402 – 2409, 2006. 1, 2

[11] M. Habbecke and L. Kobbelt. A surface-growing approach to multi-view stereo reconstruction. *CVPR*, 2007. 1, 2

[12] J.Starck, G.Miller, and A. Hilton. Volumetric stereo with silhouette and feature constraints. *BMVC*, 3. 2

[13] J.Starck and A. Hilton. Surface capture for performance based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007. 2

[14] K. Kolev, M. Klodt, T. Brox, S. Esedoglu, and D. Cremers. Continuous global optimization in multiview 3d reconstruction. *EMMCVPR 2007*, 2007. 3

[15] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. *ECCV*, pages 82–96, 2002. 1

[16] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J. M. Frahm, R. Yang, D. Nister, and M. Pollefeys. Real-time visibility-based fusion of depth maps. *ICCV*, pages 1–8, 2007. 1, 2

[17] M.Habbecke and L.Kobbelt. Iterative multi-view plane fitting. *VMV*, pages 73–80, 2006. 1

[18] M.Kazhdan, M.Bolitho, and H.Hoppe. Poisson surface reconstruction. *Eurographics Symposium on Geometry Processing*, pages 61–70, 2006. 6

[19] N.D.F.Campbell, G.Vogiatzis, C.Hernandez, and R.Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. *ECCV*, pages 766–779, 2008. 1, 2, 5

[20] N.Slesareva, A.Bruhn, and J.Weickert. Optic flow goes stereo: A variational method for estimating discontinuity-preserving dense disparity maps. *DAGM-Symposium*, pages 33–40, 2005. 2

[21] J. P. Pons, R. Keriven, and O. Faugeras. Modelling dynamic scenes by registering multi-view image sequences. *CVPR*, 2:822– 827, 2005. 1

[22] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *CVPR*, pages 519–528, 2006. 1

[23] N. Slesareva, T. Buhler, K. U. Hagenburg, J. Weickert, A. Bruhn, and H. P. Seidel. Robust variational reconstruction from multiple views. *SCIA*, pages 173–182, 2007. 3

[24] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE PAMI*, 30(6):1068–1080, 2008.

[25] T.Kanade, P.Rander, and P.Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, 4(1):34–47, 1997. 2

[26] S. Tran and L. Davis. 3d surface reconstruction using graph cuts with surface constraints. *ECCV*, pages 219–231, 2006. 1

[27] G. Vogiatzis, C. H. Esteban, P. H. S. Torr, and R. Cipolla. Multi-view stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29:2241–2246, 2007. 1, 2, 6, 7

[28] G. Vogiatzis, P. Torr, and R. Cipolla. Multi-view stereo via volumetric graph-cuts. *CVPR*, pages 391–398, 2005. 1, 2, 3, 6

[29] Y.Furukawa and J.Ponce. Accurate,dense, and robust multi-view stereopsis. *CVPR*, 2007. 1, 2, 5, 6, 7

[30] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust tv-l1 range image integration. *ICCV*, 2007. 1, 2

[31] A. Zaharescu, E. Boyer, and R. Horaud. Transformesh: A topology-adaptive mesh-based approach to surface evolution. *ACCV*, 2007. 1