Accurate Multi-view Stereopsis Fusing DAISY Descriptor and Scaled-Neighbourhood Patches

Fei Wang and Ning An(✉)

Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University,
Xi'an, Shaanxi Province, China
wfx@mail.xjtu.edu.cn, aning.393@stu.xjtu.edu.cn

Abstract. In this paper, we present an efficient patch-based multi-view stereo reconstruction approach, which is designed to reconstruct accurate, dense 3D models on high-resolution image sets. Wide-baseline matching becomes more challenging due to large perspective distortions, increased occluded areas and high curvature regions that are inevitable in MVS. Correlation window measurements, which are mainly used as photometric discrepancy function, are not appropriate for wide-baseline matching. We introduce DAISY descriptor for photo-consistency optimization of each new patch, which makes our algorithm robust on distortion, occlusion and edge regions against many other photometric constraints. Another key to the performance of Patch-based MVS is the estimation of patch normal. We estimate the initial normal of every seed patch via fitting quadrics with scaled-neighbourhood patches to handle the reconstruction of high local curvature regions. It demonstrates that our approach performs dramatically well on large-scale scene both in terms of accuracy and completeness.

Keywords: Multi-view stereo · Patch · DAISY descriptor · Normal estimation

1 Introduction

Multi-view stereopsis (MVS) is an algorithm can be described as: given a collection of images taken from an object or a scene, estimate the most likely 3D model that explains those images. It is a classic computer vision problem that occupied researchers for more than 30 years. Nowadays, more and more applications range from 3D mapping, 3D printing, virtual reality that enter our field of vision. It has seen a surge of interest that how to exploit diverse images collection ever assembled to reconstruct the 3D model about a scene [1].

We focus on the patch-based MVS algorithm and find that correlation window photometric measurements, such as NCC used in PMVS [3], are not appropriate for wide-baseline matching in MVS. Because they are not robust to perspective distortions and partial occlusions. We propose to replace NCC with DAISY descriptor [2], which let us take advantage of optimization to refine every generated patch. Another contribution is the estimation of patch normal via fitting quadrics with scaled-neighbourhood patches. It helps our approach reconstruct a scene despite the presence of occlusion or edge region. The improvement is shown on various datasets, including objects with fine

E. Chen et al. (Eds.): PCM 2016, Part I, LNCS 9916, pp. 260–270, 2016.
DOI: 10.1007/978-3-319-48890-5_26

surface details, inclined planes, deep concavities, and thin structures, outdoor scenes observed from a restricted set of viewpoints.

The rest of this paper is organized as follows: we will first review related work (Sect. 2), and provide a more detailed overview of our method (Sect. 3). We then present the individual stages of our method, including a briefly description of DAISY descriptor (Sect. 4.1), DAISY-based photometric discrepancy function (Sect. 4.2) and the estimation of patch normal (Sect. 5). Experimental results and discussions are given in Sect. 6. We conclude with results in Sect. 7.

2 Related Work

Over the last decade, the MVS problem has achieved a great development, yielding a variety of reconstruction algorithms [6]. According to the taxonomy of Seitz et al. [4], MVS algorithms can be divided into four categories: (1) 3D volumetric approaches [5] extract the surface from a 3D volume by computing a cost function. These methods usually transform 3D modelling into finding the minimum graph cut algorithm; (2) surface evolution techniques [7] include algorithm based on voxels, level sets, and surface meshes. They always demand some initialization for further optimization process, which limits their applicability; (3) algorithms that compute and merge depth maps [8, 9] perform more flexible, however, the depth maps tend to be noisy and redundant that they need steps to clean up and merge the depth maps [9]; (4) techniques that grow regions or surfaces starting from a set of extracted features [3, 11]. Furukawa [3] presented a classic algorithm, PMVS, based on patch which generates a sparse set of patches corresponding to the salient image features, and then spreads the initial matches to nearby pixels and filters incorrect matches to maintain completeness and accuracy.

In Bleyer's [12] approach, a 3D scene is represented as a collection of visually distinct and spatially coherent objects. Inspired by Markov Random Field models of image segmentation, they employed object-level color models as a soft constraint, which can improve depth estimation in powerful ways. Hoang-Hiep [13] introduced a minimum s-t cut optimization over an adaptive domain that robustly and efficiently filters a quasi-dense point cloud from outliers and reconstructs an initial surface by integrating visibility constraints, followed by a mesh-based variational refinement that captures small details, smartly handling photo-consistency, regularization, and adaptive resolution. Qi Shan [14] leveraged occluding contours to improve the performance of multi-view stereo methods. This proposed approach outperforms state of the art MVS techniques for challenging Internet datasets, yielding dramatic quality improvements both around object contours and in surface detail.

3 Algorithm Overview

Ours patch-based multi-view stereo algorithm starts from a set of calibrated images. Methods about sparse multi-view stereo and how to obtain the camera poses and intrinsic parameters please refer to VisualSFM [10].

As shown in Fig. 1, similar to the framework of PMVS, the input images are detected corner features by Difference-of-Gaussian (DoG) and Harris operators. Features are first matched across multiple images, yielding a sparse set of patches associated with salient image regions. A patch is essentially a local tangent plane approximation of a surface. Its geometry is fully determined by its center $\boldsymbol{c}(p)$, unit normal vector $\boldsymbol{n}(p)$ orient toward the cameras observing it, and a reference image $R(p)$ in which p is visible.

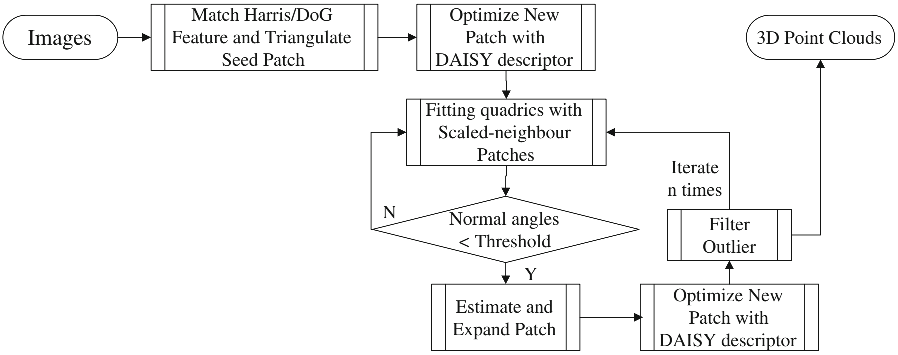
Images
Match Harris/DoG Feature and Triangulate Seed Patch
Optimize New Patch with DAISY descriptor
3D Point Clouds
Fitting quadrics with Scaled-neighbour Patches
Iterate n times
N
Normal angles < Threshold
Filter Outlier
Y
Estimate and Expand Patch
Optimize New Patch with DAISY descriptor

Fig. 1. Algorithm overview

Given a pair of features constrained to lie the corresponding epipolar lines across two images, a candidate patch p is triangulated with its center $\boldsymbol{c}(p)$, normal vector $\boldsymbol{n}(p)$, and reference image $R(p)$. Every new generated patch has its geometric parameters, $\boldsymbol{c}(p)$ and $\boldsymbol{n}(p)$, which are optimized by minimizing the discrepancy function score computed by DAISY descriptors. The patch center $\boldsymbol{c}(p)$ is constrained to lie on a ray such that its image projection in one of the visible images. $\boldsymbol{n}(p)$ is parameterized by Euler angles, yaw and pitch, yielding an optimization problem within three parameters only, which is solved by a conjugate-gradient method.

In the following step, we fuse neighbourhood information to initially estimate the patch normal. Multi-scale quadrics are fitted according to neighbourhood patches for acquiring their normals. Check if normal angles meet the threshold to determine the initial estimation of the patch normal. Expansion procedure is to spread the initial matches to nearby pixels and obtain a dense set of patches. The goal of the expansion step is to reconstruct at least one patch in every image cell. The final filter step eliminates incorrect matches and obstacles using visibility constrains. These three steps are repeated times to acquire a dense and smooth 3D point clouds.

4 Photometric Discrepancy Function

4.1 Brief Description of DAISY Descriptor

DAISY [2] is a local descriptor, inspired from SIFT and GLOH. It's combined convolved orientation maps and an isotropic Gaussian kernel that can be computed much faster. For an input image, H number of orientation maps is computed, $G_{i,}$ $1 \leq i \leq H$, where $G_o(u, v)$ equals the image gradient norm at location (u, v) for direction o. Orientation maps are written as $G_o = \left(\frac{\partial \mathrm{I}}{\partial o}\right)^+ = max(\frac{\partial \mathrm{I}}{\partial o}, 0)$, where I is the input image, o is the orientation of the derivative. Each orientation map is then convolved with Gaussian kernels of different $\sum$ values as $G_o^{\Sigma} = G_{\Sigma} * \left(\frac{\partial \mathrm{I}}{\partial o}\right)^+$.

Let $h_{\Sigma}(u, v)$ represent the vector made of the values at location (u, v) in the orientation maps after convolution by a Gaussian kernel of standard deviation $\sum$.

$$h_{\Sigma}(u, v) = \left[G_1^{\Sigma}(u, v), \ldots, G_{\mathrm{H}}^{\Sigma}(u, v)\right]^{\mathrm{T}} \tag{1}$$

The full DAISY descriptor $D(u_0, v_0)$ for location (u_0, v_0) is defined Eq. (2), where $\mathrm{I}_j(u, v, R)$ is the location with distance R from (u, v) in the direction given by j when the directions are quantized into T value.

$$\begin{aligned} D(u_0, v_0) = {} & [h^{\mathrm{T}}_{\Sigma_1}(u_0, v_0), \\ & h^{\mathrm{T}}_{\Sigma_1}(\mathrm{I}_1(u_0, v_0, R_1)), \quad \ldots, h^{\mathrm{T}}_{\Sigma_1}(\mathrm{I}_{\mathrm{T}}(u_0, v_0, R_1)), \\ & \qquad \qquad \ldots, \\ & h^{\mathrm{T}}_{\Sigma_{\mathrm{Q}}}(\mathrm{I}_1(u_0, v_0, R_{\mathrm{Q}})), \quad \ldots, h^{\mathrm{T}}_{\Sigma_{\mathrm{Q}}}(\mathrm{I}_{\mathrm{T}}(u_0, v_0, R_{\mathrm{Q}}))]^{\mathrm{T}} \end{aligned} \tag{2}$$

DAISY is parameterized with its radius R, number of rings Q, number of histograms in a ring T, and the number of bins in each histogram H. The total size of the descriptor vector is $(Q \times T + 1) \times H$.

4.2 DAISY-Based Photometric Discrepancy Function

Normalized cross correlation (NCC) is one of the most common photometric agreement measurements used in multi-view stereo algorithm. Considering it cannot work stably and robustly on the distortion, occlusion and edge region, we propose to utilize DAISY descriptor. Actually, in a worst-case scenario, DAISY will not perform any worse than a standard region-based metric like NCC. The photometric discrepancy function may not work well in the presence of specular highlights or obstacles, and we have so far assumed that the surface of a scene is nearly Lambertian as in most MVS algorithms.

Firstly, we need to determine the orientation of DAISY descriptor (Fig. 2). Given a feature, the orientation of its DAISY descriptor is defined as the vertical directions of corresponding epipolar line. It can be calculated easily. For the pair of features (f_i, f_j), the corresponding epipolar line on image i is $l_i = F^{\mathrm{T}} * f_j$, where F is the fundamental matrix. The direction vector of epipolar line $\mathbf{e} = [u, v]^{\mathrm{T}}$, rotation matrix $\boldsymbol{\theta} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, the main direction of the DAISY descriptor can be computed as

$$o = \begin{cases} \arccos(\langle \boldsymbol{\theta} \cdot \mathbf{e} \rangle / ||\mathbf{e}||), & \text{if } \boldsymbol{\theta} \cdot \mathbf{e} \cdot \mathbf{v} \geq 0 \\ \arccos(\langle \boldsymbol{\theta} \cdot \mathbf{e} \rangle / ||\mathbf{e}||) + \pi, & \text{else} \end{cases} \tag{3}$$

where **v** is a unit vector of the positive direction of vertical axis of the image.

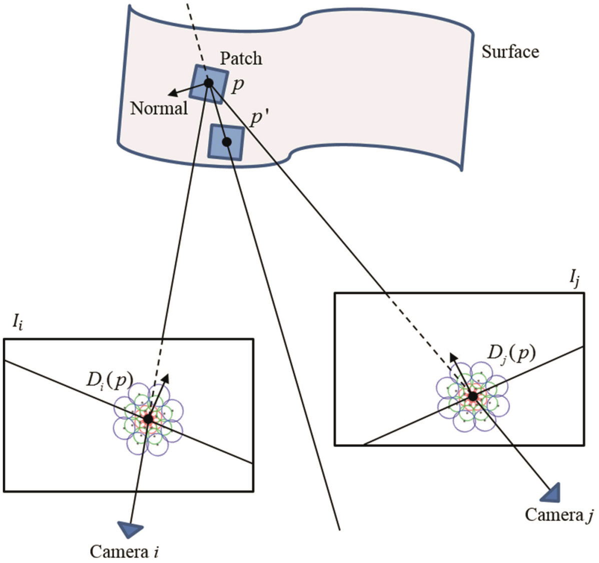

Fig. 2. Patch optimization

Now, we can get the DAISY descriptor vector $D(u, v, o)$ at the location (u, v) with the direction o. We define the dissimilarity function $d(p, I_i, I_j)$ as

$$d(p, I_i, I_j) = \frac{1}{S} \sum_{k=1}^{S} \left\| D_i^k(p) - D_j^k(p) \right\| \tag{4}$$

where $D_i(p)$ and $D_j(p)$ are the descriptors at locations obtained by projecting the patch p projected onto its visible image I_i and I_j. $D_i^k(p)$ is the kth histogram in $D_i(p)$, and S is the number of histograms used in the descriptor.

Let $V(p)$ denote a set of images in which p is visible. The photometric discrepancy function for the candidate patch p is defined as

$$c(p) = \frac{1}{|V(p) \backslash R(p)|} \sum_{I \in V(p) \backslash R(p)} d(p, I, R(p)) \tag{5}$$

We advocate replacing correlation window measurements with local region DAISY descriptor. In more complicated scenes, where images have varying resolutions and their location is non-uniformly, it becomes critical to adjust the size of the domain.

To improve the robustness of our approach, only images whose pairwise photometric discrepancy score with the reference image $R(p)$ is below a certain threshold α are used for further estimation. And we replace $V(p)$ in the photometric discrepancy function (5) with $V^*(p)$ to obtain the new formula (7). If the number of visible images of the candidate patch $|V^{*}(p)|$ is smaller than the threshold γ, it is failed and retry the next candidate patch ($\alpha = 0.6$, $\gamma = 3$). Otherwise, a new seed patch is generated successfully.

$$V * (p) = \{I | I \in V(p),\ d(p, I, R(p)) \le \alpha\} \tag{6}$$

$$c * (p) = \frac{1}{|V * (p) \backslash R(p)|} \sum_{I \in V*(p) \backslash R(p)} d(p, I, R(p)) \tag{7}$$

5 Normal Estimation

It is difficult to reconstruct a 3D model with PMVS where the local curvature is too high, including deep concavities and high convexities, because each new patch is constrained to lie on a ray for optimization such that its image projection in one of the visible images. From Fig. 3, the dashed lines are the initial estimation of candidate patch normal in PMVS. It can be assumed almost true when the reconstructed surface exactly faces towards cameras, while some high curvature regions are unavoidable. We propose to fuse the multi-scale neighbourhood information around the candidate patch to acquire a more accurate initial estimation of normal.

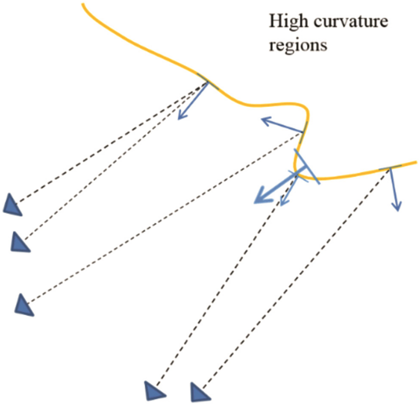

Fig. 3. Cameras and the surface of a scene

We search $\sigma \times K$ neighbourhood patches to fit a quadric and search $\lambda\sigma \times K$ neighbourhood patches to fit another quadric in a lower scale ($K \in [30,60]$, $\sigma = 2$, $\lambda = 0.6$ in our experiment). Then we get two normals n_1, n_2. When the two normals satisfy the formula (8), the normals are regarded as the same direction, and the normal computed in higher scale is set to be the initial estimation of the new patch. Otherwise, $\lambda^2\sigma \times \mathrm{K}$ neighbourhood patches are searched to compute the normal n_3, checking whether n_2 and n_3 meet the threshold β ($\beta = \pi/9$). And repeat the above process. If the two normals still have comparatively large deviation, the normal computed in smaller scale is used to as the initial estimation of the new patch. With multi-scale neighbourhood information, the generated patch normal can be adjusted more effectively no matter on the plain regions or high curvature regions.

Right now the new seed patch already has a relatively accurate initial normal, and its geometric parameters, $\boldsymbol{c}(p)$ and $\boldsymbol{n}(p)$, can be refined by the further optimization of our DAISY-based photometric discrepancy function.

$$\cos^{-1}(normal_1, normal_2) < \beta \tag{8}$$

6 Results

Our algorithm is implemented by VC++ with the CGAL library. All experiments are conducted on a Windows PC with Intel i5 CPU @3.3 GHz, RAM 16 GB.

We compare and analyze our 3D reconstruction results with that of PMVS [3] method on four datasets. The datasets are acquired from the real-world scenes. We have one dataset taken from a sculpture by ourselves, and other three are the open datasets from Stecha et al. [15]. When we reconstruct a scene or an object, it is firstly captured form various viewpoints, then the structure-from-motion software [10] is used to reconstruct the pose of each camera and obtain the projection matrix for per visible image. We take images and their corresponding projection matrix as the input of the further patch-based multi-view stereopsis approaches.

The first dataset "*Typography-P9*", its image resolution and number of images are listed in Table 1, followed by the parameters of DAISY descriptor and *K*-neighbourhood numbers. This dataset has its surface composed of many inclined planes, and parts of them have characters carved with fine details. From the comparison of 3D models (Fig. 4), our DAISY-based method performs obviously better than PMVS on the completeness. The regions with red circles appear large holes because these areas are inclined and close to the edge region. Our DAISY descriptor photometric discrepancy

Table 1. Information of datasets and parameters set in experiments

Datasets	Resolution	Images	*R*	*Q*	*H*	*T*	*K*
Typography-P9	3264 × 2448	9	5	3	8	8	40
Castle-P10	3072 × 2048	10	5	3	8	8	50
Fountain-P11	3072 × 2048	11	5	3	8	8	50
Herz-Jesu-P8	3072 × 2048	8	5	3	8	8	50

Table 2. Experimental data

Method	*Typography-P9*	*Castle-P10*	*Fountain-P11*	*Herz-Jesu-P8*
PMVS	355142	1572701	1348187	1236046
DAISY-based	481073	1724573	1621941	1438512
Final method	486249	1849036	1644479	1537345
Percent	36.92 %	17.57 %	21.98 %	24.38 %

function performs robust against NCC used in PMVS. Table 2 gives the comparison of numbers of patches. And the denseness of our reconstructions improve a lot.

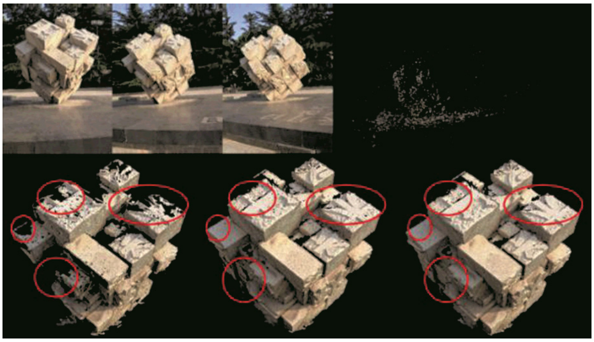

Fig. 4. *Typography-P9*. Top row: images from the dataset, the seed patches. Bottom row: from left to right, the 3D models reconstructed by PMVS, our DAISY-based and final method (Our MVS method with DAISY-based measurement and normal estimation).

On the datasets from Stecha et al. [15], our DAISY-based method and final method are proved to reconstruct more smooth 3D models (Fig. 5). *Castle-P10* has an obvious character of bending building wall, which is more apparently showed the limitation of PMVS. It is almost impossible to rebuild by PMVS on the region of bending building

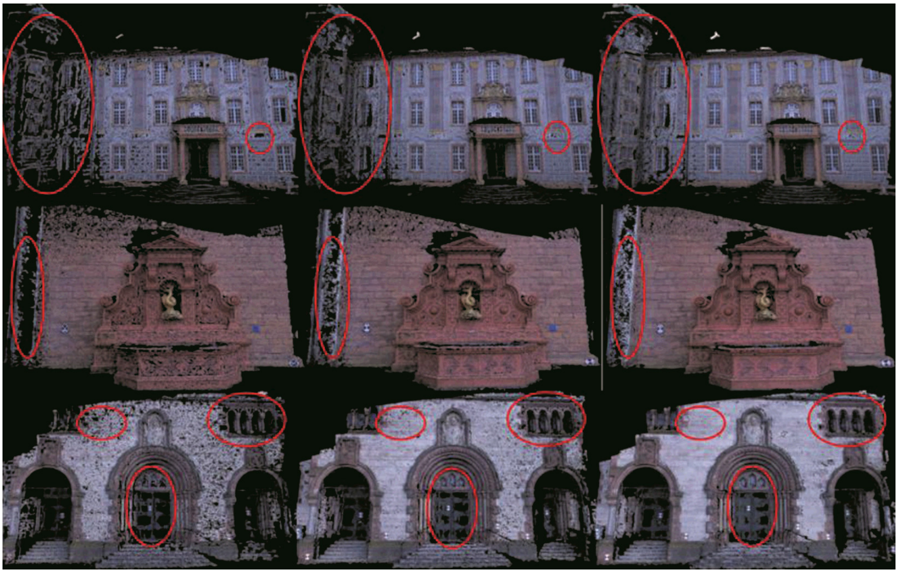

Fig. 5. Datasets from Strecha et al. [15]. From top to bottom row, they are *Castle-P10, Fountain-P11, Herz-Jesu-P8.* And from left to right column, they are 3D models reconstructed by PMVS, our DAISY-based method and our final method.

wall, whereas our method has an estimation of patch normal fusing scale-neighbourhood information which helps perform well. There are building walls and some deep concavities, and thin structures on the dataset *Fountain-P11*. And our method produces more dense patches than that of PMVS with fine surface details (more clearly shown in Fig. 6). *Herz-Jesu-P8* has a large plain region, which leads to the fish-like scale of the 3D model reconstructed by PMVS with small correlation window. However, large size may lead to holes on the thin structure regions because unique local regions are weakly supported. It is necessary to replace correlation window measurements with stable local region descriptor for it's critical to adjust the size of the window.

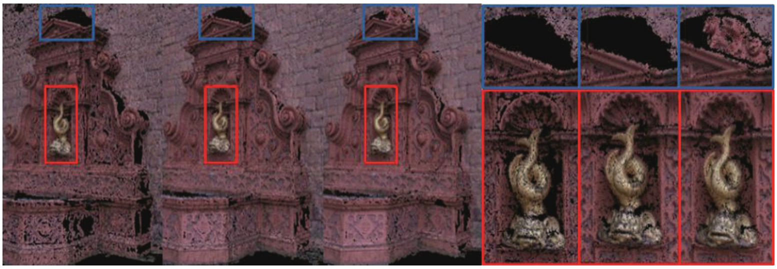

Fig. 6. Details of 3D models reconstructed by PMVS, our DAISY-based and final method.

Some details are given in Fig. 6, and it is obviously showed that our methods can produce better reconstructed results on the surface details, edge regions, deep concavities, occlusion areas, and thin structures. To measure quantitative evaluations of the reconstructed 3D models, we have two 3D point clouds comparing with their ground truths in Fig. 7. The correct rate is measured by changing the deviation of 3D point clouds from their corresponding actual depth value after 3D point clouds are aligned with their ground truth by iterative closest point (ICP).

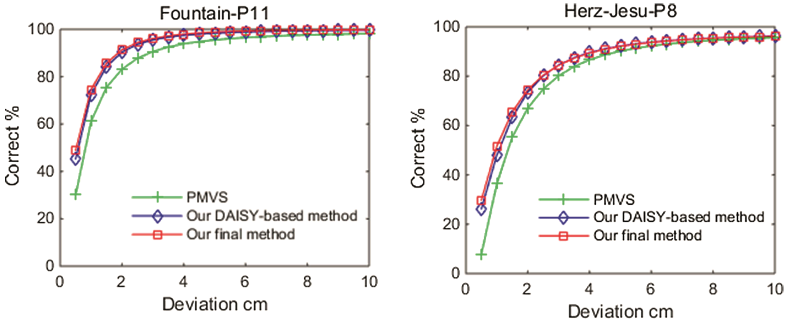

Fig. 7. Quantitative evaluation

7 Conclusion

In this paper, we introduced DAISY descriptor into patch-based MVS to be as photometric discrepancy function for wide-baseline matching, which performs robust than correlation window measurements. Especially on areas of different depths, edge regions or partial occlusions, PMVS cannot reconstruct for perspective distortions. Our estimation of patch normal via fusing scaled-neighbourhood information enhances the sensibility to the region of high curvature. It contributes to handle the areas of deep concavities and high convexities. The experiments are proved that the method proposed by us has a better performance on the high curvature regions, slanted surfaces and thin structures. In comparison to PMVS, the 3D models reconstructed by our algorithm have a great improvement on the accuracy and completeness.

Acknowledgment. National Natural Science Foundation of China (No. 61231018, No. 61273366), National Science and technology support program (2015BAH31F01), Program of introducing talents of discipline to university under grant B13043.

References

1. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. Int. J. Comput. Vis. **80**(2), 189–210 (2008)
2. Tola, E., et al.: Daisy: an efficient dense descriptor applied to wide-baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 815–830 (2010)
3. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE Trans. Pattern Anal. Mach. Intell. **32**(8), 1362–1376 (2010)
4. Seitz, S.M., Curless, B., Diebel, J., et al.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 519–528 (2006)
5. Sormann, M., Zach, C., Bauer, J., Karner, K., Bishof, H.: Watertight multi-view reconstruction based on volumetric graph-cuts. In: Ersbøll, B.K., Pedersen, K.S. (eds.) SCIA 2007. LNCS, vol. 4522, pp. 393–402. Springer, Heidelberg (2007). doi:10.1007/978-3-540-73040-8_40
6. Furukawa, Y., Hernández, C.: Multi-view stereo: a tutorial. Found. Trends Comput. Graph. Vis. **9**(1–2), 1–148 (2015)
7. Esteban, C.H., Schmitt, F.: Silhouette and stereo fusion for 3D object modeling. Comput. Vis. Image Underst. **96**(3), 367–392 (2004)
8. Xiao, X., et al.: Multi-view stereo matching based on self-adaptive patch and image grouping for multiple unmanned aerial vehicle imagery. Remote Sensing **8**, 89 (2016)
9. Shen, S., Hu, Z.: How to select good neighboring images in depth-map merging based 3D modeling. IEEE Trans. Image Process. **23**(1), 308–318 (2014)
10. Wu, C.: Towards linear-time incremental structure from motion. In: 2013, International Conference on 3D Vision-3DV, pp. 127–134. IEEE (2013)
11. Zhu, Z., Stamatopoulos, C., Fraser, C.S.: Accurate and occlusion-robust multi-view stereo. ISPRS J. Photogrammetry Remote Sens. **109**, 47–61 (2015)
12. Bleyer, M., Rother, C., Kohli, P., et al.: Object stereo - Joint stereo matching and object segmentation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3081–3088 (2011)

13. Vu, H.H., Labatut, P., Pons, J.P., et al.: High accuracy and visibility-consistent dense multiview stereo. IEEE Trans. Pattern Anal. Mach. Intell. **34**(5), 889–901 (2011)
14. Qi, S., Curless, B., Furukawa, Y., et al.: Occluding contours for multi-view stereo. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 4002–4009 (2014)
15. Strecha, C., von Hansen, W., Gool, L.V., et al.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8 (2008)