Refinement of Surface Mesh for Accurate Multi-View Reconstruction

Radim Tyleček and Radim Šára
{tylecr1,sara}@cmp.felk.cvut.cz

Center for Machine Perception
Faculty of Electrical Engineering,
Czech Technical University,
Prague, Czech Republic

Abstract. In this paper we propose a pipeline for accurate 3D reconstruction from multiple images that deals with some of the possible sources of inaccuracy present in the input data. Namely, we address the problem of inaccurate camera calibration by including a method [1] adjusting the camera parameters in a global structure-and-motion problem which is solved with a depth map representation that is suitable to large scenes.

Secondly, we take the triangular mesh and calibration improved by the global method in the first phase to refine the surface both geometrically and radiometrically. Here we propose surface energy which combines photo consistency with contour matching and minimize it with a gradient method. Our main contribution lies in effective computation of the gradient that naturally balances weight between regularizing and data terms by employing scale space approach to find the correct local minimum. The results are demonstrated on standard high-resolution datasets and a complex outdoor scene.

1 Introduction

The development of methods for 3D reconstruction from multiple images has led to a number of successful methods, belonging to the group of multi-view stereo (MVS) algorithms [2–5]. Despite the effort and availability of high resolution images, their performance is still not satisfying when we compare them to the laser range measurement systems [6]. The fact that high resolution images can be easily obtained by consumer cameras or downloaded from the web is motivation for improving the results of MVS algorithms, also when the time and hardware costs of range scanning are considered.

Traditionally, evaluation is performed in the terms of *completeness* and *accuracy* [7]. When comparing these two criteria, completeness is more related to models that help resolve ambiguities present in stereo matching. On the other hand, accuracy relates more to the chosen representation and an ability to handle deviations of the input data from the model. Keeping in mind that these two views still share a wide base, we will focus on the second one in this paper, and propose a pipeline to deal with some of the possible sources of inaccuracy in MVS.

2 Mesh refinement pipeline

Our idea is first to use a global method [1] to improve calibration and possibly to obtain inaccurate estimate of the surface, represented as a set of depth maps, which is followed by change of representation to mesh that allows a local approach to variational correction of vertices.

The camera parameters originate typically from sparse correspondences [8] and can be further improved with dense data by one of the methods [2, 1]. Both methods demonstrate that camera calibration update is easily tractable with image based (depth maps or patches) representations, while other are not suitable for this purpose, i.e. with volumetric or mesh representations the update would be difficult.

While the depth map representation in image space is useful for large scenes and natural to the input data, it has limits for modeling arbitrary surfaces as it is not intrinsic to them. A change of representation is thus required for further improvement of the surface accuracy. The global method [1] provides us with a good initial estimate of the surface, represented by a discrete triangular mesh, and a refined camera calibration. We choose this mesh as a suitable discrete representation that is intrinsic to the surface, and denote it as a set of vertices $\mathbf{X}_i \in \mathbb{R}^3, i = 1, \ldots, n_X$ and triangle indices $\mathbf{T}_j \in \{1, \ldots, m\}^3, j = 1, \ldots, n_T$.

For the purpose of deriving our method, we will start with continuous definition, and later discretize the results. In this task, our goal will be to find the estimate of surface S by the minimization of a surface energy E_ϕ:

$$E_\phi(S) = \int_S \phi(\mathbf{X}) dA, \tag{1}$$

where $\phi(\mathbf{X})$ is a photo-consistency measure and dA is surface element. Since we assume a good initial estimate of the surface S, we can resort in our method to implicit regularization of the surface based on the minimal surface area.

The primary goal in multi-view reconstruction is to find a surface with photo consistent projections to multiple images. Traditionally, MVS algorithms [9] measure the photo consistency of a given surface point with views observing it from normal direction, or choosing views close to it. This comes from the fact that deviation from a Lambertian model becomes critical when observing surface under large angle with respect to the surface normal. Additionally, we can also exploit the information from views observing the surface from tangential direction, what leads to contour matching.

In the following sections we will combine these two sources to construct ϕ and next propose a method for its minimization.

2.1 Photo-consistency measure

We define a photo-consistency function ϕ_I for a given world point $\mathbf{X}$ and a set of images $I_i, i = 1, \ldots, N$ in the following way:

$$\phi_I(\mathbf{X}) = \sum_{i,j \in V(\mathbf{X}),\, i \neq j} \frac{2\|I_i(\pi_i(\mathbf{X})) - I_j(\pi_j(\mathbf{X}))\|^2}{\sigma_i^2(\pi_i(\mathbf{X})) + \sigma_j^2(\pi_j(\mathbf{X}))} \tag{2}$$

where $V(\mathbf{X})$ is a set of images in which point $\mathbf{X}$ is visible, and $\pi_i(\mathbf{X}) \simeq \mathbf{P}_i\mathbf{X}$ is perspective projection function ($\mathbf{P}_i$ is a camera matrix). The normalizing factors $\sigma_{i,j}$ are independently pre-computed variances of image functions $I_{i,j}$ in visible regions and they estimate expected measurement error assuming a Poisson distribution of the image values. They allow the range of the measuring function to be approximated by $\phi \in \langle 0, 1+\epsilon \rangle, \epsilon \ll 1$. Our resulting measure is thus a *normalized sum of squared differences* (NSSD). As pointed out in [3], we avoid the use of normalized cross correlation (NCC), which introduces additional ambiguities.

The traditional Lambertian assumption allows us to use simple difference of pixel intensities, unfortunately this model is often violated, for instance, the exposure parameters are often different in available input images. Since modeling of reflectance properties is complex, i.e. with radiance tensors [10], we will limit ourselves to intensity offset correction. We will thus attempt to find the 'true' offset-corrected images $I_i^* = I_i - C_i$ which minimize the total error (2) by choosing the offset C_i to be the mean radiance error of the surface visible in camera i:

$$C_i = \frac{1}{N_i} \sum_{j \mid i \in V(\mathbf{X}_j)}^{N_i} \Big(I_i(\pi_i(\mathbf{X}_j)) - \bar{I}(\mathbf{X}_j) \Big), \tag{3}$$

where $\bar{I}(\mathbf{X})$ is the mean of the projections of point $\mathbf{X}$ to images where it is visible, being the best estimate of radiance with respect to square error in (2), and N_i is the number of vertices X visible in camera i. Now we can replace original images I_i with corrected I_i^* in all our image terms derived from (2).

2.2 Contour matching

The analysis of [11] has first brought the observation that projection of contour generators on a smooth surface should match local maxima of image gradient ∇I (apparent contours), which has recently been an inspiration for [12, 13]. Similarly to [13] we avoid explicit detection of contours in images by a more general formulation, but we additionally take into account the directions of ∇I and surface normals $\mathbf{N}$ projected to the image. It be formalized by maximization of a contour matching function $\phi_C(\mathbf{X})$:

$$\phi_C(\mathbf{X}) = \frac{1}{|\Omega(\mathbf{X})|} \sum_{k \in \Omega(\mathbf{X})} \Big| \Big\langle \nabla I\big(\pi_k(\mathbf{X})\big), \varpi_k\big(\mathbf{N}(\mathbf{X})\big) \Big\rangle \Big|, \tag{4}$$

where $\varpi_k\big(\mathbf{N}(\mathbf{X})\big) = \frac{\pi_k(\mathbf{N}(\mathbf{X}))}{\|\pi_k(\mathbf{N}(\mathbf{X}))\|}$ is a unit normal projected to the image and $\langle \cdot, \cdot \rangle$ is a scalar product. We denote here $\Omega(\mathbf{X})$ as the set of cameras that see $\mathbf{X}$ as a contour point. Inversely, for a given camera k, we can find contours Ω_k on the surface S as curves, where normal $\mathbf{N}(\mathbf{X})$ of each of its visible points is perpendicular to the viewing direction $\mathbf{X} - \mathbf{C}_k$:

$$\Omega_k = \big\{ \mathbf{X} \,|\, \big\langle \mathbf{N}(\mathbf{X}), \mathbf{X} - \mathbf{C}_k \big\rangle = 0, k \in V(\mathbf{X}) \big\}, \tag{5}$$

where $\mathbf{C}_k$ is the camera center. In practice for discrete meshes, we identify contour vertices by change of the sign of the dot product above and change of visibility. Now

we can partition surface points in the following sets for every camera k: V_k – set of points visible in camera k, $\bar{V}_k$ – set of points not visible in camera k and Ω_k – points generating contour in camera k.

To adapt our method for large datasets, we limit the size of V_k by choosing only a given number of the best views based on the angle between the normal and view direction, calculated from the dot product in (5).

We can now put together photometric and contour measures in

$$E_\Omega(S) = \int_S \Big(\phi_I(\mathbf{X}) - \alpha\phi_C(\mathbf{X})\Big) dA = \int_S \phi(\mathbf{X}) dA, \tag{6}$$

where $\phi_I(\mathbf{X})$ is integrated in cameras $k \in V(\mathbf{X})$ and $\phi_C(\mathbf{X})$ in $k \in \Omega(\mathbf{X})$. Parameter α allows control of the preference between contour and image matching; we used $\alpha = 1$ in our experiments.

2.3 Gradient-based approach

According to [14, p. 22], we can obtain a surface flow that minimizes the energy (1) by

$$\frac{\partial S}{\partial t}(\mathbf{X}) = \Big(H(\mathbf{X})\phi(\mathbf{X}) - \langle \nabla\phi(\mathbf{X}), \mathbf{N}\rangle\Big)\mathbf{N}, \tag{7}$$

where $H(\mathbf{X})$ is the mean curvature of surface at point $\mathbf{X}$. The solution S^* is found by Euler's time integration of (7), hence deforming the surface by

$$\mathbf{X}_{t+dt} = \mathbf{X}_t + dt\,\frac{\partial S}{\partial t}(\mathbf{X}_t), \tag{8}$$

where dt is a chosen time step used in iterations.

The first part of the flow (7) performs implicit regularization, for $\phi(\mathbf{X}) \to 1$ this flow corresponds to *mean curvature flow*, which leads to minimization of the surface area. In our flow this applies to areas with high photometric error, and on the other hand, for low error $\phi(\mathbf{X}) \to 0$ has no effect. This kind of balancing between regularization and data gets around the shrinking effects of pure surface minimization present in many

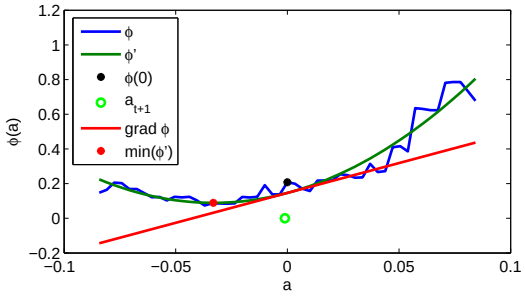


Fig. 1. Photo-consistency $\phi(\tilde{\mathbf{X}}(a))$ sampled in the normal direction with curve ϕ' fitted to it. Value of $a = 0$ corresponds to the current vertex position.

variational methods. The second part of (7) moves the surface along normal $\mathbf{N}(\mathbf{X})$ in the direction where $E(S)$ will decrease, which can be calculated by taking the negative projection of the gradient to the normal movement direction. For regions with missing data (vertices $\mathbf{X}_0$ visible in less than two views), the minimal surface should be the optimal solution, which is accomplished by setting $\phi(\mathbf{X}_0) = 1$.

We compute the directional derivative $\langle \nabla\phi(\mathbf{X}), \mathbf{N} \rangle$ by sampling points $\tilde{\mathbf{X}}(a), a \in \langle -\tau, \tau \rangle$ along the normal in images I^* for $k \in V(\mathbf{X})$ or in the image gradient ∇I^* for $k \in \Omega(\mathbf{X})$ and computing $\phi(\tilde{\mathbf{X}}(a))$, like in Figure 1. At this point we discretize the problem by counting the energy integral (6) only in the vertices $\mathbf{X}_i$ of the mesh, so the photo consistency is evaluated in individual mesh vertices and no image neighborhood is used. We use this simplification efficiently with mesh resampled so that the mean of edge projection to images is around 2-3 pixels.

In order to avoid falling to a local minimum, the derivative is computed from a quadratic polynomial $\phi'(\tilde{\mathbf{X}}(a)) = p_1 a^2 + p_2 a + p_3$ fit to the samples. In order to perform with a limited number of samples, the window specified by τ is gradually decreased in iterations: $\tau_t = \tau_0 \gamma^{t-1}$ where t is iteration, $\gamma = 0.95$ is the decrease rate and τ_0 is initial window size determined from average edge sizes around vertex $\mathbf{X}$. This means that in the first iterations the decision is based on wider support and allows us to find a global minimum in the initial window. In later iterations the region near this minimum is sampled more densely, producing a more precise estimate.

This can also be thought of as regularizing data with a scale determined by the window size. When computing a gradient from the initial large window, the curve cannot fit the data exactly and is rather flat, resulting in a smaller gradient and more smoothing. The data weight is increased as the window size decreases, when the fitted curve gets steeper and the gradient size is higher. Window size control is more natural than explicitly adjusting the second term in (7) with a constant increasing over iterations: if there is no strong minimum (i.e. in noisy conditions), the gradient will not increase and the model will not over fit here.

3 Experiments

We have evaluated our method first on four high accuracy datasets from a publicly available benchmark [6], which allows comparison of the results with a number of other state-of-the-art methods both in quantitative and qualitative ways, by analyzing occupancy histograms and diffuse renderings. The original results of the depth map fusion [1] were taken as the input for the mesh refinement procedure. In all cases, the algorithm was run for 30 iterations, when the window size τ drops to 20% of the initial size.

The quantitative evaluation in [6] was performed with ground truth acquired with time-of-flight laser measurement. Evaluated scene is projected to the input cameras and obtained depths are compared with the ground truth depths in the scale of their accuracy σ, which is shown in Figure 2 for *fountain-P11* dataset. More results are available on the benchmarking website[1]. The results of refinement show relative increase of accuracy

[1] http://cvlab.epfl.ch/ strecha/multiview/denseMVS.html

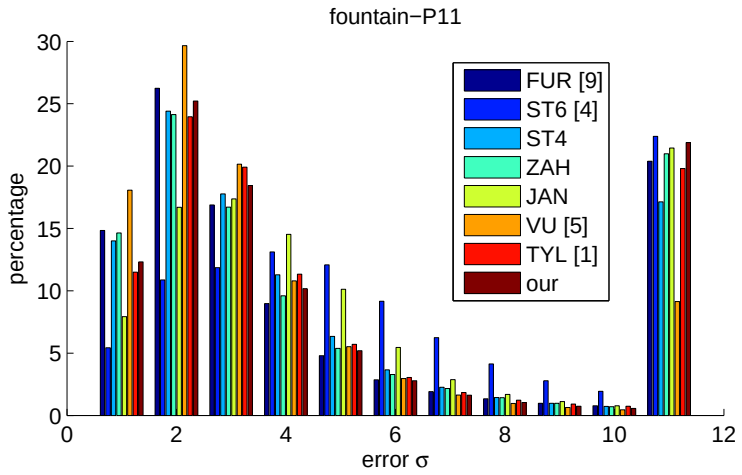
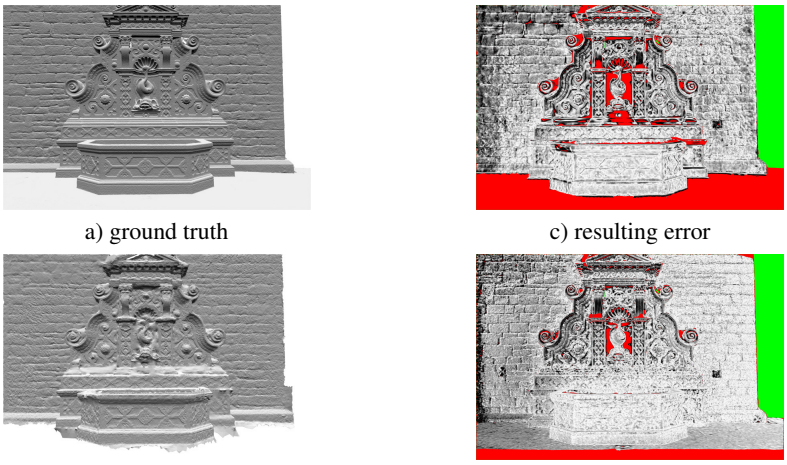


Fig. 2. Histogram from [6], each bin represents percentage of pixels with an error equal to $n\sigma$. Accuracy manifests by higher value in bins $n = 1, 2$.

from initial depth map fusion output by 7.2% at $\sigma \leq 2$. Use of this score for direct comparison of accuracy with other methods is difficult, since we are here evaluating our surface very close to the accuracy limit of the ground truth (σ is the measurement variance). Also the result depends substantially on the completeness of the surface, i.e. the currently best scoring method [5], which combines the best of several previous methods, succeeds in reconstructing the ground plane of *fountain-P11*, what adds to all bins of the histogram in Figure 2. Still, they miss the camera calibration adjustment in their pipeline, and thanks to this feature our method is able to achieve higher accuracy in certain areas, like in Figure 4 g), h) and i), while the error is distributed evenly over the surface in Figure 3 c).

However, this quantitative evaluation does not take into account the quality of the surface. Although estimated surface may be close to the ground truth, the human observer is influenced by regularity or smoothness of the surface, i.e. when resulting 3D models are used for visualization. For this purpose, comparison of surface normals would be appropriate, but while it is not included in [6], we will use the renderings in its place. Figure 4 presents results in this way and shows how the initial result of depth map fusion in c) was improved by the refinement in d) with flat surfaces are smoothed and edges emphasized. Here similar results of the best performing state-of-the-art methods [9, 5] in e) and f) still show notable roughness.

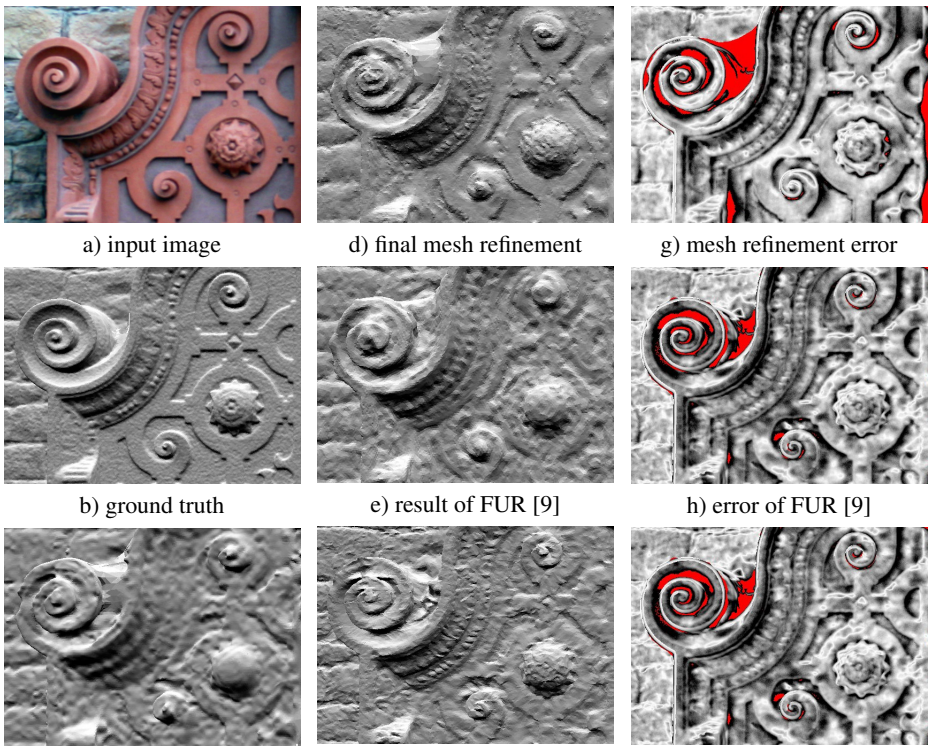
In order to evaluate the effect of individual contributions to the accuracy of the proposed method, we have run it with different modifications on the *fountain-P11* dataset. The results can be compared visually in detail in Figure 5. The importance of the contour matching term is demonstrated on the difference between a) and d), where the edges become bumpy. It can be also seen from this comparison that the majority of the edges are recognized as contour generators (ϕ_C), including the sunken ornaments, after they are first 'discovered' by image matching (ϕ_I). On the other hand, we can encounter



a) ground truth c) resulting error

b) our result d) error of VU [5]

Fig. 3. *Fountain-P11* dataset [6] overview diffuse rendering and error maps. Accurate regions are white, missing reconstruction is red and green area was not evaluated.

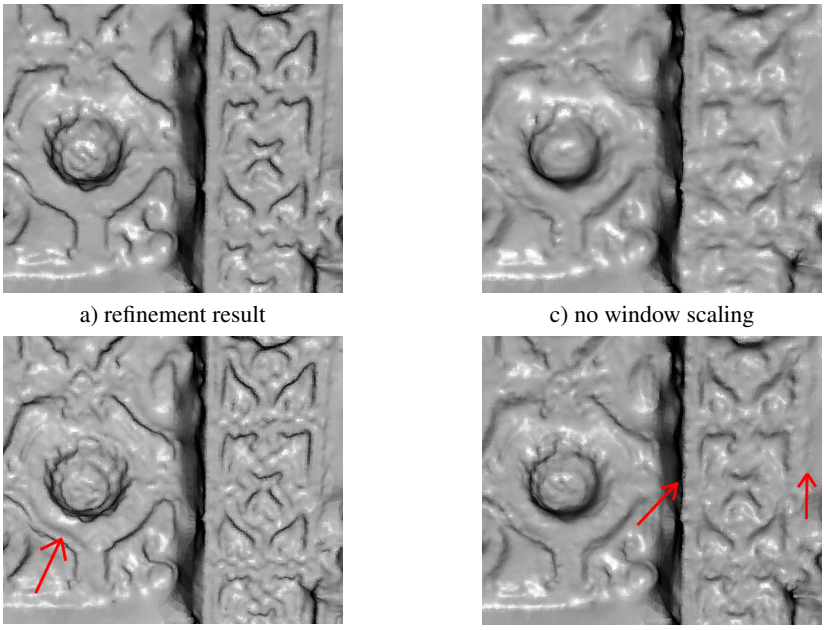


a) input image d) final mesh refinement g) mesh refinement error

b) ground truth e) result of FUR [9] h) error of FUR [9]

c) initial depth map fusion f) result of VU [5] i) error of VU [5]

Fig. 4. *Fountain-P11* dataset [6] detailed rendering and error maps (white=accurate, black=inaccurate,red=missing).



a) refinement result

c) no window scaling

b) no offset correction

d) no contour term

Fig. 5. Demonstration of effect of individual contributions on *Fountain-P11* dataset [6].

false contour generators detected on noisy initial surface, which can cause the surface to create phantom edges. This has particularly effect on textured surfaces, and it has to be avoided by more robust detection of contour generators. Next, without image offset correction in b), surface in flat regions becomes sinuous while the edges are correct thanks to the contour information as it is invariant to image offset errors. Finally, when we omit the iterative scale space approach in c), the surface becomes globally oversmoothed or eventually overfitted to data depending on the fixed window size.

To demonstrate the possibilities of the method on large scale data, we have used it to reconstruct the statue *Asia*, which is a part of the Albert Memorial in London. We captured a suite of 238 photographs (Figure 6), which consists of several semi-rings, three monocular from about 2m, 4m and 40m distance and one stereo with non-uniform (free-hand) vertical baseline from about 8m distance plus some additional images. All photos have been shot by Canon PowerShot G7 (10 Mpix) with variable focal length and with image stabilization on, and carefully corrected for radial distortion. The variable lighting conditions (moving clouds) were compensated by our offset correction (up to 25% of the intensity range). The model reconstructed with depth map fusion [1] shown in Figure 7 includes intricate features like elephant's tusks, but some parts of the surface are only approximated due to missing data (tops and some back parts of the statue). We performed subsequent refinement in the same way as previous datasets. Since we have no ground truth data available, the effect of refinement can be demonstrated visually by introduction of details, like the rug on the elephant's head in Figure 8 c).

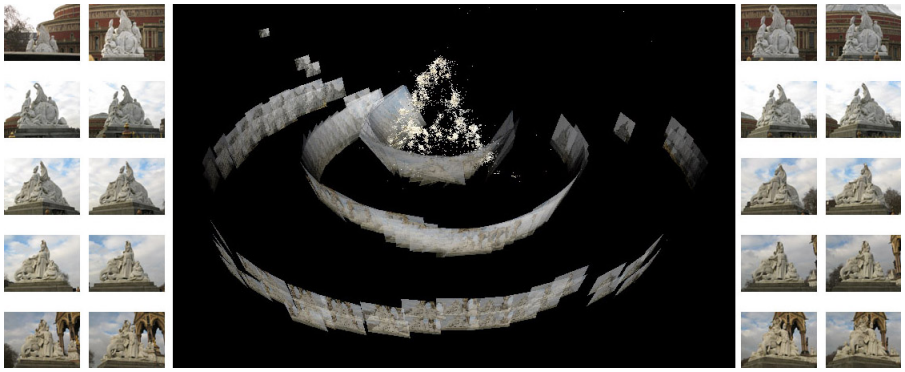


Fig. 6. *Asia* dataset scene (Albert Memorial, London) with sparse points and some of 238 input images.

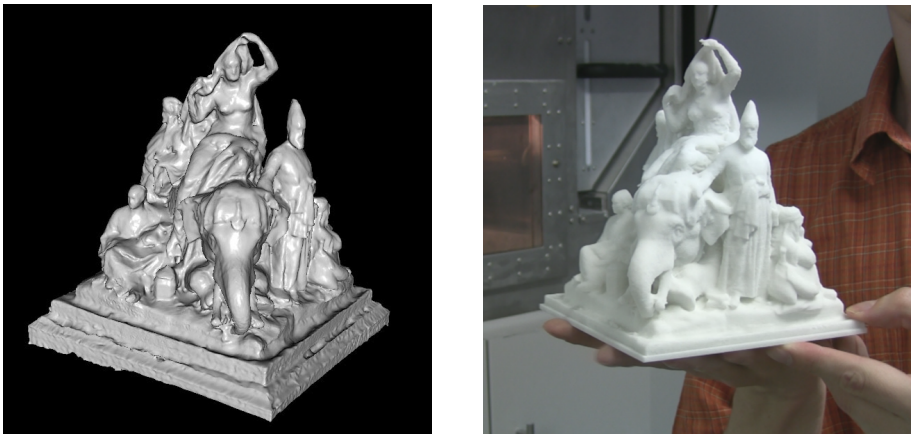
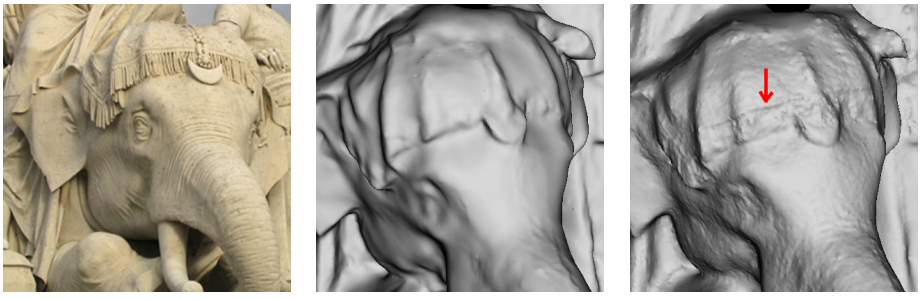


Fig. 7. Results on the *Asia* dataset. Left: initial model produced with depth map fusion [1]. Right: replica produced with rapid prototyping from the final model refined by our method.



a) input image b) initial surface c) refined surface

Fig. 8. Demonstration of mesh refinement on the *Asia* dataset, elephant's head in detail.

4 Conclusion

We have proposed a method towards increasing accuracy in MVS. Variable 3D surface representation allows us to achieve efficient camera pose refinement together with surface geometry refinement. Surface contour modeling helps utilize independent sources of 3D shape information present in the images, while image offset correction compensates for the effect of their exposure and scale-space approach is employed to find the correct surface within noisy data. In our future work we plan tying the processes of calibration and refinement more closely together.

Acknowledgments

This work was supported by CTU Prague project CTU0912713 and by EC project FP6-IST-027113 eTRIMS.

References

1. Tyleček, R., Šára, R.: Depth Map Fusion with Camera Position Refinement. In: Proc Computer Vision Winter Workshop, Eibiswald, Austria (February 2009) 59–66
2. Furukawa, Y., Ponce, J., Team, W.: Accurate Camera Calibration from Multi-View Stereo and Bundle Adjustment. In: Proc CVPR. (2008) 8
3. Goesele, M., Snavely, N., Curless, B., Hoppe, H., Seitz, S.: Multi-view stereo for community photo collections. In: Proc ICCV. (2007)
4. Strecha, C., Fransens, R., Van Gool, L.: Combined Depth and Outlier Estimation in Multi-View Stereo. In: Proc CVPR. (2006) 2394–2401
5. Vu, H., Keriven, R., Labatut, P., Pons, J.P.: Towards high-resolution large-scale multi-view stereo. In: Proc CVPR. (June 2009)
6. Strecha, C., von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On Benchmarking Camera Calibration and Multi-View Stereo for High Resolution Imagery. In: Proc CVPR. (2008)
7. Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In: Proc CVPR. (2006) 519–528
8. Martinec, D., Pajdla, T.: Robust rotation and translation estimation. In: Proc CVPR. (June 2007)
9. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multi-view stereopsis. In: Proc CVPR. (2007)
10. Soatto, S., Yezzi, A.J., Jin, H.: Tales of shape and radiance in multi-view stereo. In: Proc ICCV. (2003) 974–981
11. Koenderink, J.: What does the occluding contour tell us about solid shape. Perception **13**(3) (1984) 321–30
12. Delaunoy, A., Prados, E., Gargallo, P., Pons, J., Sturm, P.: Minimizing the Multi-view Stereo Reprojection Error for Triangular Surface Meshes. In: Proc BMVC. (2008)
13. Keriven, R.: A variational framework for shape from contours. Technical report, Ecole Nationale des Ponts et Chaussees, CERMICS, France (2002)
14. Jin, H.: Variational methods for shape reconstruction in computer vision. PhD thesis, Washington University, St. Louis, MO, USA (2003)