Watertight Multi-view Reconstruction Based on Volumetric Graph-Cuts

Mario Sormann[1], Christopher Zach[1], Joachim Bauer[1], Konrad Karner[1], and Horst Bishof[2]

[1] VRVis Research Center,
Inffeldgasse 16, 8010 Graz, Austria
sormann@vrvis.at
[2] Institute for Computer Graphics and Vision, Graz University of Technology
Inffeldgasse 16, 8010 Graz, Austria
bischof@icg.tu-graz.ac.at

Abstract. This paper proposes a fast 3D reconstruction approach for efficiently generating watertight 3D models from multiple short baseline views. Our method is based on the combination of a GPU-based plane-sweep approach, to compute individual dense depth maps and a subsequent robust volumetric depth map integration technique. Basically, the dense depth map values are transformed to a volumetric grid, which are further embedded in a graph structure. The edge weights of the graph are derived from the dense depth map values and if available, from sparse 3D information. The final optimized surface is obtained as a min-cut/max-flow solution of the weighted graph. We demonstrate the robustness and accuracy of our proposed approach on several real world data sets.

Keywords: volumetric 3D reconstruction, graph-cut, dense depth maps, virtual 3D models.

1 Introduction

In our approach we consider the problem of creating virtual 3D models solely from a set of digital input images, which is still a challenging problem in computer vision. The principal reason for utilizing digital images as input source, is the independancy of the 3D reconstruction process from the size of the objects to be modeled.

Current state of the art approaches for multi-view reconstruction are divided in two main categories: one pass (or directed) methods versus two pass (or indirect) methods. Direct methods, recently proposed by Vogiatzis et al. [19] or Hornung and Kobbelt [9] process all available input images from different viewpoints simultaneously. Their methods are based on finding a minimum cut in a graph structure, which is embedded in a volumetric grid. One of the main benefits of these methods is that they generate watertight surfaces such that the final 3D model does not contain any disturbing holes. Clearly, a drawback is that these approaches still rely on existing object silhouettes to consider only voxels which are close to the visual hull. But, the extraction of visual hull information, especially for complex environments, can be a tedious and time consuming process. Therefore we introduce an indirect, two pass method which extracts in a

B.K. Ersbøll and K.S. Pedersen (Eds.): SCIA 2007, LNCS 4522, pp. 393–402, 2007.

first pass a set of dense depth maps, whereas the second pass enforces a robust integration of the depth maps to create proper and watertight 3D models. Thus, we are able to provide intermediate results and can bring, if necessary, a human operator into the reconstruction loop. Especially for large data sets, an user assisted visual evaluation of intermediate results can be very helpful to detect errors at the earliest possible point in the 3D reconstruction pipeline. Therefore, we try to combine the main benefits of both, direct and indirect approaches. Consequently, the main contributions of our approach are the following:

1. Our approach avoids the incorporation of visual hull information, because the extraction of visual hull information is a tedious and time consuming process.
2. As a side effect of our indirect reconstruction process, we can easily bring a human operator into the reconstruction loop for quality assessment.
3. The proposed method is able to reconstruct 3D models even from dense depth maps containing outliers.
4. Due to the fact that our method utilizes global minimization techniques we can guarantee a watertight and global optimized surface.
5. Our algorithm can deal with high volumetric resolutions as well as a large number of input images.

2 Related Work

The automatic 3D reconstruction of complex objects is still an active research field within the computer vision community. There are two major approaches to the problem of 3D real world modeling: range-based modeling and image-based modeling. Range-based modeling is based on laser scanners. A very well known approach in this field is The Digital Michelangelo Project carried out by M. Levoy et.al. [15].

In this work we focus on image-based modeling, which represents the 3D reconstruction of real world objects from a dense set of photographs. A comparative evaluation of image-based and range-based methods can be found in El-Hakim and Beraldin [5]. Image-based modeling techniques utilize in general widely available hardware and developed systems can be used for a wide range of different objects and scenes. Furthermore such algorithms produce realistic models with an increasing level of automation.

All range-based methods as well as most of the image-based modeling methods generate 2.5D heightfields. In order to generate true 3D models, a robust fusion of this set of heightfields into a single 3D surface is necessary. The fundamentals of robust depth image fusion in the context of laser scanned data was proposed by Curless and Levoy [3]. The basic idea of volumetric range image integration is the conversion of depth maps to corresponding 3D distance fields and a subsequent robust averaging of these distance fields. The resolution and the accuracy of the final model are determined by the quality of the source images and the resolution of the target volume. Recently, Zach et. al. [23] introduced a fast GPU-based method, based on the original work of Curless and Levoy [3]. Since this method is a pure local method, the final 3D model can still contain many holes.

In contrast shape from silhouette methods try to overcome these restrictions. They recover the shape of the objects from their contours, known as visual hull, and no depth

map information is used. A practical system to generate 3D models from its profiles was introduced by Wong and Cipolla [20]. This approach uses only the silhouettes of a sculpture for both motion estimation and model reconstruction, and neither corner detection nor matching is necessary. The method is robust and fast, but as drawback they are limited to simple shaped objects. Therefore, recent developed methods combine the visual hull information with a photo-consistency function, which is further embedded in a graph structure. A general approach combining multi-camera stereo reconstruction with graph-cuts was presented by Kolmogorov et. al. [13]. A comparison of energy functional types, which can be minimized using graph-cuts is given by Kolmogorov and Zabih in [12]. Several applications of graph-cut based energy minimization for volumetric reconstruction were presented in Vogiatzis et. al. [19] and Hornung and Kobbelt [9]. In these approaches, individual voxels correspond to nodes in the graph, used to determine the maximum flow. These techniques still rely on existing object silhouettes in order to consider only voxels close to the visual hull. Additionally, visibility information is mainly introduced from the visual hull to find occluded views for each voxel.

The inspiration for the method presented in this paper is given by a number of above mentioned volumetric 3D reconstruction approaches and efficient energy minimization techniques utilizing graph-cuts. More precisely, most of the above mentioned methods have in common that it is in general difficult to generate watertight and global optimized 3D models from dense depth maps, which is the standard output of most image-based modeling techniques. Furthermore, discussed methods either perform the 3D reconstruction in two passes, but then can not guarantee a watertight and global optimized surface, or in one pass, but then bypass the dense depth maps and extract the 3D model directly. Consequently, there is still a need to combine the ideas and benefits of both schema.

3 Dense Depth Map Estimation

Our work targets the reconstruction of objects from arbitrary image sequences taken with a calibrated digital consumer camera. The process of camera calibration and pose estimation, which are not the topics of this paper, are well studied problems in computer vision and determine the internal and external parameters of a camera [8].

The set of images with known calibration and orientation is used to generate a 3D model of the object in a fully automated manner. For dense depth map estimation a fast reconstruction method suitable for small-baseline settings is applied for every view. Basically, we utilize a plane-sweep approach [21] to create the set of dense depth maps, using up to 5 images simultaneously for matching (one key image in the middle and one or two neighboring reference images on each side). For each depth value, the reference images are projected onto the key image plane, located at the given depth and a correlation measure with respect to the key image is calculated. Occlusion handling is addressed by the best half-sequence strategy. The set of slices filled with correlation values comprise a data structure similar to the disparity space image. A final matching algorithm (e.g. scanline optimization [17]) establishes the dense depth map from the disparity space image. Depending on the resolution, plane-sweep matching requires 5.5 seconds for each reference image at an resolution of 1024x1024 pixels. More details of our developed GPU-based plane-sweeping technique can be found in Zach et. al. [23].

4 Graph-Cut Based Volumetric Depth Map Integration

In this section we give an overview of the basic ideas, datastructures and processing steps of our approach. All different steps are discussed in more detail in the following subsections. An overview is given in Figure 1.

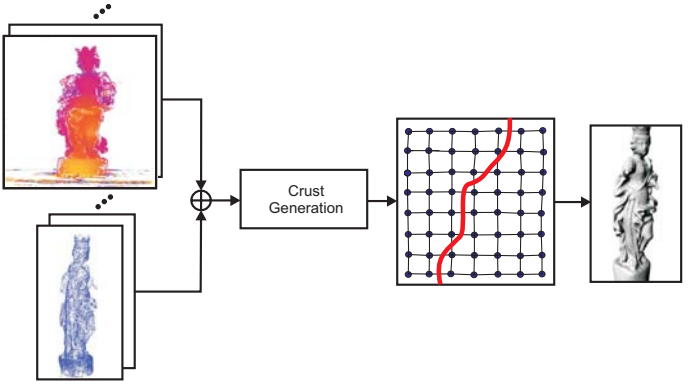


Fig. 1. Overview of our graph-cut based volumetric depth map integration pipeline: First, dense deph map values and if available sparse 3D information is transformed to a volumetric grid. Then, we direct extract the surface confidence in the vicinity of the dense depth maps. Finally, a min-cut/max-flow algorithm is performed to determine a watertight global optimized surface on the selected volumetric resolution. Note, that all available 3D information are in same coordinate system.

The required input for the volumetric integration approach is a set of dense depth maps and, but not necessarily, sparse 3D information obtained beforehand as proposed by Bauer et. al. [1]. Instead of utilizing visual hull information as proposed by Vogiatzis et. al. [19], we derive a so called crust band directly in the vicinity of the dense depth maps. This crust band can be interpreted as a confidence map, which represents the probability that the final unknown surface passes through. The confidence values are computed as an unsigned distance function ϕ over the underlying volumetric grid, which is described in more detail in section 4.1.

As soon as the confidence values are computed, we determine a global optimized surface S_{opt}, which approximates the true but unknown surface, with respect to the used energy functional. Previous work already have shown that such problems can be efficiently solved by a min-cut/max-flow algorithm [19]. Additionally, we incorporate sparse 3D information into our energy functional, which further enhance the obtained 3D reconstruction results.

Finally, the voxel based representation is transformed into a triangular mesh based on a standard marching cube algorithm introduced by Lorenson and Cline [16].

4.1 Crust Generation

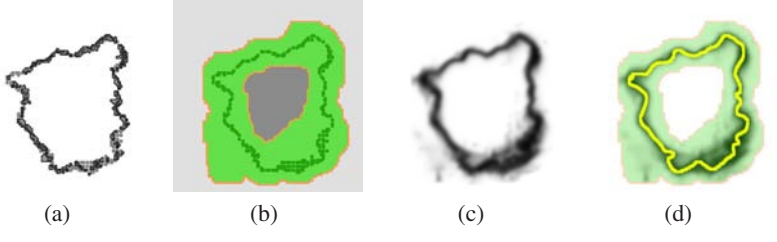
The first step in our approach is the determination of crust voxels lying on both sides of the true surface. The generated crust should be as small as possible in order to obtain the

maximal computational efficiency. On the other hand, the crust must be able to reflect potential concavities arising in the true model geometry. Consequently, the generation of the proper crust is non-trivial, and recently proposed strategies include incorporation of the visual hull [19] and coarse-to-fine approaches [10].

We select a different path by employing the initial, still noisy 3D result of our efficient depth map integration scheme as the primary indicator of crust voxels. Our volumetric depth image integration method [23] robustly averages the set of approximated signed distance fields induced by the depth maps. For every voxel a statistic is accumulated, which is based on the signed distance of the voxel to the approximately closest surface point indicated by the current depth map. Finally, a voting scheme determines the final signed distance value of a voxel, which can be used to extract the isosurface. We utilize the accumulated signed distance field to determine the initial set of crust voxels by including voxels close to the isosurface (with respect to a user-specified distance threshold). This set is enhanced by a number of dilation steps d to achieve a watertight separation of interior and exterior regions. In all our experiments we generally set $d = 2$.

Since isosurfaces generated from signed distance tend to have unnecessary high genus, positive surface confidences $\phi(v)$ are employed in the extraction procedure instead of signed distance, using a similar approach to [10]. Voxels crossed by the isosurface as well as voxels which are filled from sparse 3D information have confidence value zero (indicating high certainty), and the confidences of all other crust voxels are initialized with 1. The confidence map ϕ is subsequently smoothed using a homogeneous diffusion scheme.

Figure 2 illustrates all intermediate results of our initial crust generation process.



(a) (b) (c) (d)

Fig. 2. This image illustrates all intermediate results showing one slice of the volumetric grid of the St. Barbara data set. **(a)** Dense depth map values (light grey) and sparse 3D information (dark grey). **(b)** Obtained voxel crust (green), exterior (light grey) and interior (dark grey) component **(c)** Confidence band derived from dense depth map values, where darker values correspond to higher confidence. **(d)** Optimal surface (blue) extracted by a min-cut/max-flow algorithm.

4.2 Surface Reconstruction

This section is dedicated to discuss our graph-cut based surface reconstruction procedure. Since, our goal is to extract an optimal as well as watertight surface S_{opt}, we transform the volumetric grid to a graph based structure and solve the optimization problem by performing a min-cut/max-flow algorithm. Similar to other approaches we minimize $E(D) = \sum_{v \in D} \omega(v)$ where D is a weighted sum of dense depth map values.

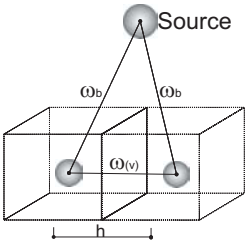


Fig. 3. Correspondence between crust voxels and nodes in the graph

The geometric configuration of the correspondence between crust voxels and nodes in the graph is shown in Figure 3.

The corresponding graph, consisting of all crust voxels, is connected over a regular six-neighborhood. The edge weight $\omega(v)$ is assigned to all edges of the embedded graph and can be derived from the unsigned distance value $\phi(v)$. Basically, $\omega(v)$ is defined as $\omega(v) = (\phi(v))^s$, where s can be interpreted as some kind of smoothness factor. Additionally, as discussed in Vogiatzis et. al. [19] we add a ballooning force ω_b, which connects every crust voxel to the source node with a constant weight of $\omega_b = \lambda h^3$, where λ is a weight parameter and and h represents the quantized size of a voxel. The ballooning force avoids a cut across thin structures of the object. As usual, interior voxels V_{int} are connected to the source node and exterior voxels V_{ext} are connected to the sink node. As stated, for exploiting sparse 3D information we extend $\omega(v)$ in the following way:

$$\omega(v) = \begin{cases} (\phi(v))^s & \forall v \in D_V \\ 0.0 & \forall v \in E_V \end{cases} \tag{1}$$

where D_V and E_V represents dense depth map values and sparse 3D information respectively.

After the min-cut/max-flow algorithm has determined the optimal surface voxels S_{opt}, a standard marching cube algorithm converts the voxel based surface into a triangular mesh for possible further processing.

To summarize, our approach reconstructs watertight 3D surface models, even from non-outlier free dense depth maps. In contrast to related approaches and due to the fact that our approach do not rely on visual hull information we avoid the complex, time consuming and tedious task of acquiring such information. In addition, we do not need any hole filling algorithm, since large gaps are effectively closed due to the embedded energy functional. And finally, incorporated sparse 3D information enhances the quality of our final 3D models.

5 Results

This section is dedicated to discuss the visual and quantitative results of our approach. We applied our depth map integration method to several real-world data sets. All experiments were performed on 4 GHZ PC with 2GB main memory and a GeForce 7800 GT

with 256MB graphics memory. The images were taken with a calibrated digital consumer camera at a geometric resolution of 4064x2704 pixels. After pose estimation, the source views are resized to 1024x1024 pixels and the obtained dense depth map have the same resolution (unless noted otherwise).

Table 1 demonstrates quantitative results and compares the number of input images, target resolution, mesh complexity as well as the timing for each of our data sets. The reconstruction time includes the dense depth map estimation as well as the volumetric depth map integration, which is the less dominant computational factor.

Table 1. Illustration of time and space complexity for each of our data sets. The obtained reconstruction time can be separated into a dense depth map estimation part and a volumetric depth map integration part, which is the more dominant computational factor.

Dataset	Images	Resolution	Triangles	Time [min.]
Barbara	46	256x384x256	704446	9.5
Pedestal	74	256x256x384	890358	14.5
Temple	47	256x256x384	790186	7.5

The first data set depicted in Figure 4 shows the lime stone statue of St. Barbara, which was reconstructed from 46 images. The statue is 55cm tall with a diameter of 13cm at the pedestal. The final 3D model was reconstructed in less than 10 minutes and consists of approximately 700k triangles.

(a) (b) (c) (d) (e)

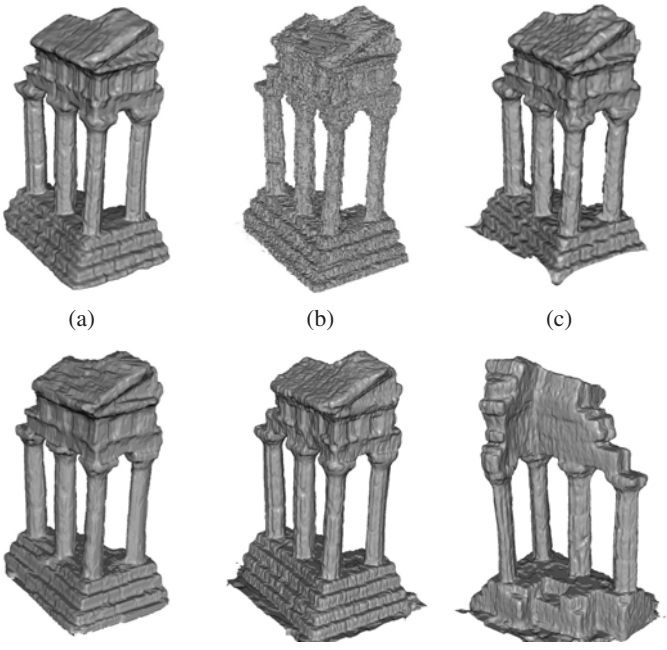
Fig. 4. 3D Reconstruction of the statue of St. Barbara from 46 images. **(a)** One input image of the data set. **(b)** Obtained dense depth map. **(c)-(d)** Two viewpoints of the reconstructed 3D model consisting of approximately 700k triangles. **(e)** Textured version of our obtained 3D reconstruction.

The second experiment (Figure 5) illustrates a pedestal (3x2x1.5 meters) of a statue located in front of the Austrian National Library in Vienna. We obtained the final 3D model in about 15 minutes at an geometric resolution of 900k triangles. For the reconstruction of the pedestal we used 76 images. We are able to obtain a visually appealing as well as watertight 3D model, even in textureless regions around the fresco's.

Finally, Figure 6 illustrates the well known temple data set from the multi-view stereo evaluation page [18] consisting of 47 images. The dense depth maps were obtained

(a) (b) (c) (d) (e)

Fig. 5. 3D Reconstruction of a pedestal, located in front of the Austrian National Library from 76 images. **(a)** One image of the data set. **(b)** Obtained dense depth map. **(c)-(d)** Two viewpoints of the final 3D reconstruction consisting of approximately 900k triangles. **(e)** Textured 3D model.



(a) (b) (c)

(d) (e) (f)

Fig. 6. 3D reconstruction of the well known temple data set from the multi-view stereo evaluation page [18] consisting of 47 input images. **(a)** 3D reconstruction proposed by Furukawa and Ponce [6]. **(b)** 3D reconstruction of Kolmogorov and Zabih [11]. **(c)** Obtained 3D model of Vogiatzis et. al. [19]. **(d)** 3D reconstruction of Hornung and Kobbelt [9]. **(e-f)** Two views of our achieved 3D reconstruction consisting of approximately 800k triangles.

at a resolution of 640x480 pixels. The visual comparison of our results against four other related multi-view reconstruction methods is shown in Figure 6(a-d). Figure 6(e-f) illustrates two views of our 3D reconstruction consisting approximately 800k triangles. Note, that all presented results, except the one shown in Figure 6(a), are utilizing graph-cuts for global optimization. Of course, the quantitative as well as qualitative evaluation of our results is given at the multi-view stereo evaluation page [18].

6 Conclusion

In this paper we demonstrated a fast and robust method for the 3D reconstruction of proper 3D models, even from non-outlier free dense depth maps. The achieved quality of our 3D models mainly depends on the grade of the dense depth maps as well as the selected target resolution. One main advantage of the proposed method is, that there is no need for some kind of visual hull information during the 3D reconstruction process. Due to a min-cut/max-flow optimization we can guarantee a watertight and global optimized surface.

Though the results are very promising there are several improvements that can made to our approach. Further work needs to include the already generated error map, which provides a confidence measurement for each dense depth map value, into the cost functional of the min-cut/max-flow algorithm. Finally, we plan to evaluate and compare several edge weight functions.

Acknowledgments

This work is partly funded by the VRVis Research Center, Graz and Vienna/Austria (`http://www.vrvis.at`). We would also like to thank the Vienna Science and Technology Fund (WWTF).

References

1. Bauer, J., Zach, C., Karner, K., Bischof, H.: Efficient sparse 3d reconstruction by space sweeping. In: 3DPVT (Chapel Hill, USA, June 2006). CD Proceedings (2006)
2. Collins, R.T.: A space-sweep approach to true multi-image matching. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (San Francisco, USA, June 1996), pp. 358–363 (1996)
3. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: ACM SIGGRAPH (New Orleans, USA, August 1996), vol. 1, pp. 303–312 (1996)
4. Debevec, P.E., Taylor, J., Malik, J.: Modeling and rendering architecture from photographs. In: ACM SIGGRAPH (New Orleans, USA, August 1996), pp. 11–20 (1996)
5. El-Hakim, S., Beraldin, J.: Configuration analysis for sensor integration. In: Proceedings of SPIE (Philadelphia, USA, October 1995), vol. 2, pp. 274–285 (1995)
6. Furukawa, Y., Ponce, J.: High-fidelity image-based modeling. Tech. rep. UIUC (2006)
7. Goesele, M., Curless, B., Seitz, S.: Multi-view stereo revisited. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (New York, USA, June 2006), vol. 1, pp. 2402–2409 (2006)
8. Heikkilä, J.: Geometric camera calibration using cicular control points. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 10 (October 2000), pp. 1066–1077 (2000)
9. Hornung, A., Kobbelt, L.: Hierarchical volumetric multi-view stereo reconstruction of manifold surfaces based on dual graph embedding. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (New York, USA, June 2006), vol. 1, pp. 503–510 (2006)
10. Hornung, A., Kobbelt, L.: Robust reconstruction of watertight 3d models from non-uniformly sampled point clouds without normal information. In: Eurographics Symposium on Geometry Processing (Sardinia, Italy, June 2006), vol. 1, pp. 41–50 (2006)

11. Kolmogorov, V., Zabih, R.: Multi-camera scene reconstruction via graph cuts. In: European Conference on Computer Vision (Copenhagen, Denmark, May 2002), vol. 3, pp. 82–96 (2002)
12. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 2 (February 2004), 147–159 (2004)
13. Kolmogorov, V., Zabih, R., Gortler, S.J.: Generalized multi-camera scene reconstruction using graph cuts. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Madison, USA, June 2003), vol. 1, pp. 501–516 (2003)
14. Kutulakos, K.N., Seitz, S.M.: A theory of shape by space carving. International Journal of Computer Vision 38 3, 199–218 (2000)
15. Levoy, M., Pulli, K., Curless, B., Rusinkiewicz, S., Koller, D., Pereira, L., Ginzton, M., Anderson, S., Davis, J., Ginsberg, J., Shade, J., Fulk, D.: The digital michelangelo project: 3d scanning of large statues. In: ACM SIGGRAPH (New Orleans, USA, July 2000), vol. 1, pp. 131–144 (2000)
16. Lorenson, W., Cline, H.: A high resolution 3d surface reconstruction algorithm. In: ACM SIGGRAPH (Anaheim, USA, July 1987), vol. 1, pp. 163–170 (1987)
17. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 47(1-3), 7–42 (2002)
18. Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (New York, USA, June 2006), vol. 1, pp. 519–526 (2006)
19. Vogiatzis, G., Torr, P.H.S., Cipolla, R.: Multi-view stereo via volumetric graph-cuts. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Washington D.C., USA, June 2005), vol. 1, pp. 391–398 (2005)
20. Wong, K.-Y., Cipolla, R.: Reconstruction of outdoor sculptures from silhouettes under approximate circular motion of an uncalibrated hand-held camera. IEEE Transactions on Information and Systems 87(1), 27–33 (2004)
21. Yang, R., Welch, G., Bishop, G.: Real-time consensus based scene reconstruction using commodity graphics hardware. In: Proceedings of Pacific Graphics (Beijing, China, October 2002), pp. 358–363 (2002)
22. Yezzi, A., Soatto, S.: Stereoscopic segmentation. International Journal of Computer Vision 53(1), 31–43 (2003)
23. Zach, C., Sormann, M., Karner, K.: High performance multi-view reconstruction. In: 3DPVT (Chapel Hill, USA, June 2006). CD Proceedings (2006)