A PASSIVE 3D SCANNER
Acquiring High-quality Textured 3D-models Using a Consumer Digital-camera

Matthias Elter, Andreas Ernst and Christian Küblbeck
Fraunhofer Institute for Integrated Circuits (IIS), Am Wolfsmantel 33, 91058 Erlangen, Germany

Keywords: 3D scanner, passive 3D reconstruction, shape from stereo, structure from motion, texture mapping, volumetric fusion, dense stereo.

Abstract: We present a low-cost, passive 3d scanning system using an off-the-shelf consumer digital camera for image acquisition. We have developed a state of the art structure from motion algorithm for camera pose estimation and a fast shape from stereo approach for shape reconstruction. We use a volumetric approach to fuse partial shape reconstructions and a texture mapping technique for appearance recovery. We extend the state of the art by applying modifications of standard computer vision techniques to images of very high resolution to generate high quality textured 3d models. Our reconstruction results are robust and visually convincing.

1 INTRODUCTION

Acquiring geometric models of physical objects using active scanning techniques is a solved problem. A great variety of mature 3d scanning techniques (based on laser triangulation or structured light) is available today. However, all of these techniques have the major drawback that they acquire an object using active sensors. Furthermore the required hardware is usually expensive. A less intrusive and cheaper approach is to reconstruct both the shape and appearance (color and texture) of an object from images acquired using a standard digital camera. A passive sensor like a standard digital camera is both cheap and does not change the object which is to be acquired (for example by illuminating it). In this paper we present a passive 3d scanning technique which reconstructs both 3d shape and appearance of arbitrary objects from 2d images acquired by a consumer digital camera. To achieve this we have developed a reconstruction approach based on standard computer vision concepts like structure from motion and shape from stereo. We extend the state of the art by applying modifications of these techniques to images of very high resolution to generate high quality models and textures.

2 STATE OF THE ART

Many approaches to passive 3D shape reconstruction can be found in literature. The most important basic techniques that are employed are *shape from stereo*, *shape from silhouette*, *shape from shading*, and *shape from focus*. Shape from stereo techniques reconstruct 3d shape from point correspondences in two or more images. A taxonomy and evaluation of shape from stereo algorithms can be found in a recent survey (Scharstein and Szeliski, 2002). Shape from silhouette algorithms reconstruct the 3d shape of an object from a sequence of 2d silhouette (contour) images of the object. Laurentini (Laurentini, 1994) and Kutulakos (Kutulakos and Seitz, 1998) provide a theoretical foundation of the concepts exploited by shape from silhouette algorithms. Implementations include (Tarini et al., 2002) and (Andrew W. Fitzgibbon, 1998). Shape from shading approaches try to estimate shape from the shading pattern (light and shadows) of in a single image of an object. Here shape is estimated in the sense of a field of normals from which a surface can be recovered up to scale. Zhang and Tsai evaluate six well-known shape from shading algorithms in a recent survey paper (Zhang et al., 1999). A technique that tries to estimate object shape by changing the camera intrinsics is shape from fo-

Elter M., Ernst A. and Küblbeck C. (2007).
A PASSIVE 3D SCANNER - Acquiring High-quality Textured 3D-models Using a Consumer Digital-camera.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IU/MTSV*, pages 311-316

cus. Image pixels corresponding to different depths, obviously, will be optimal in focus for different settings of the camera intrinsics like the focal length. Examples of shape from focus approaches are (Ziou, 1998), (Schechner and Kiryati, 2000) and (Favaro and Soatto, 2002).

Much less publications on approaches that describe complete passive 3d scanning systems, including shape and appearance recovery, can be found in literature. Examples are (Wolfgang Niem, 1997), (Weik, 2000) and (Andrew W. Fitzgibbon, 1998).

3 METHODS

3.1 Image Acquisition and Camera Calibration

Images are acquired using a cheap (300$) Panasonic DMC-FZ3 consumer digital camera. We make use of its continuous drive mode, which allows to continuously take three frames per second at full resolution of 2015×1512 pixels. With the camera in continuous drive mode mounted on a tripod, an image sequence, showing the object that is to be scanned from multiple viewpoints, is acquired by rotating it in front of the camera. The intrinsic camera parameters are obtained by our robust extension (Rupp et al., 2006) of Zhangs classic camera calibration technique (Zhang, 1998). We furthermore use the lens distortion coefficients obtained by the camera calibration to remove lens distortion effects from the images.

3.2 Structure from Motion

For camera pose estimation we have implemented a state of the art structure from motion algorithm. Feature points are detected and tracked from view to view using the approaches introduced by Kanade, Shi and Tomasi (Tomasi and Kanade, 1991; Shi and Tomasi, 1994). Due to occlusion and because of features leaving the field of view of the camera, points usually can not be tracked throughout all views of the image sequence. Therefore lost features are constantly replaced by newly detected points while processing the image sequence. We have implemented a sequential structure from motion algorithm: first an initial pair of views and initial structure is reconstructed. Then additional views and additional structure are sequentially added to the initial reconstruction. Figures 1, 2 and 3 illustrate the sequential approach of our algorithm.

Figure 1: Initial views (gray) and structure (blue) of a human head scene. View positions and orientations are illustrated using pyramid glyphs.

3.2.1 Initial Structure and Motion

The fundamental matrix $\mathbf{F}$ for the initial pair of views v_1 and v_2 is estimated using a robust estimator. We have developed a genetic algorithm approach similar to the one described by Rodehorst (Rodehorst, 2004) instead of using the classic RANSAC algorithm. We are using a calibrated camera and hence can obtain the essential matrix $\mathbf{E}$ directly from $\mathbf{F}$. We then obtain the metric camera projection matrices $\mathbf{P_1}$ and $\mathbf{P_2}$ corresponding to the two initial views from the $\mathbf{E}$ by means of factorization (Hartley and Zisserman, 2003). Using $\mathbf{P_1}$ and $\mathbf{P_2}$, initial structure is then obtained from the corresponding feature points of the initial view pair by means of triangulation.

3.2.2 Adding Views

Pose and structure of the remaining views of the image sequence are now sequentially added to the reconstruction. Based on 3D/2D point correspondences between the already reconstructed structure and feature points in a new view v_n, its camera projection matrix $\mathbf{P_n}$ can be estimated. Again, we are using a genetic algorithm as a robust estimator. Already reconstructed structure is then refined and new points are added by means of triangulation. For robust triangulation of feature points that are visible in more than the minimal two views, we use a robust least-median-of-squares based estimator. Once all views of the image sequence are added to the reconstruction, it is refined by a global optimization step using bundle adjustment (Brown, 1976).

3.3 Shape from Stereo

Structure from motion as described above results in a sparse reconstruction (point cloud) of an object. We however are interested in a dense reconstruction. Hence we use shape from stereo concepts to obtain

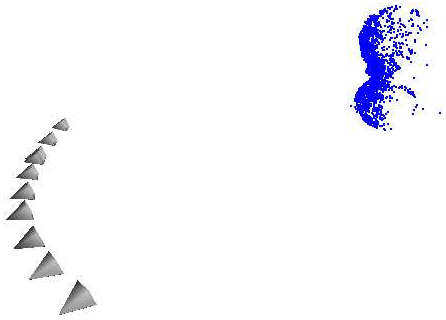

Figure 2: The human head scene after adding seven additional views.

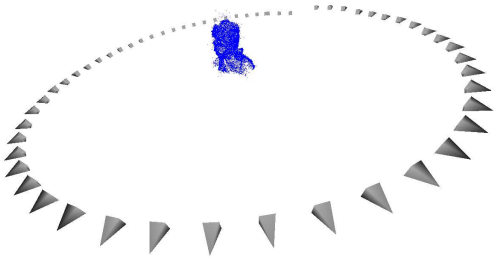

Figure 3: The full structure from motion reconstruction of the human head scene. The structure (blue) is discarded and the views (gray) are used for the following shape from stereo step.

dense reconstructions from pairs of close views of the image sequence. We then merge these 2-view reconstructions using a volumetric fusion approach to obtain the desired full reconstruction of an object.

3.3.1 Obtaining 2-View Reconstructions

We reconstruct partial reconstructions from two views each using a fast template matching algorithm. We apply two preprocessing steps to reduce the complexity of the matching problem and to improve the robustness of the dense stereo matcher. We reduce the correspondence search space by one dimension using nonlinear image rectification (Oram, 2001). The rectified images are then transfered from color space to census space using the census transform (Zabih and Woodfill, 1994). The census transform is a nonparametric local transformation that relies on the relative ordering of local intensity values instead of the intensity values them self. Our own experiments and a recent survey (Brown et al., 2003) indicate that matching in census space increases the robustness

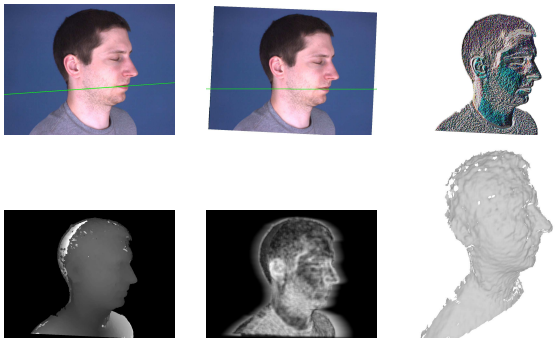

Figure 4: Source images (top left) are rectified (top middle) and the census transform is applied (top right). The green line illustrates an epipolar line which is equivalent with a scanline after rectification. Using template matching a disparity map (bottom left) and a confidence map (bottom middle) are obtained. Finally a triangle mesh is created (bottom right).

against lighting differences and occlusions. We have developed a template matching based dense stereo matcher. We match in census space where the standard difference metrics like normalized cross correlation can not be used. Instead we use the Hamming distance. Given the very high resolution of our input images and the fact that we work on three color channels, acceptable CPU complexity for the dense matcher can only be achieved by avoiding redundant computations. Hence we have implemented a very fast computation scheme that avoids redundant computations (Stefano and Mattoccia, 2000). The resulting disparity map contains errors, which are due to texture-less areas, repetitive patterns and occlusions. We enforce the uniqueness and the left-right stereo constraints to identify and remove most of the erroneous areas in the disparity maps. As the integer accuracy of the block matcher would lead to step effects in the reconstruction and hence is not good enough for a smooth reconstruction we use a postprocessing step to achieve subpixel accuracy (Frischholz, 1997). For the volumetric fusion, we are also interested in the confidence of the disparity values. Hence we define the difference of the minimal matching difference and the mean matching difference as a confidence metric. Finally we obtain a triangle mesh from the disparity map by means of triangulation. Figure 4 illustrates the individual steps of our 2-view reconstruction algorithm.

3.3.2 Volumetric Fusion

The 2-view reconstructions are now fused using a volumetric approach. The triangle meshes are transfered

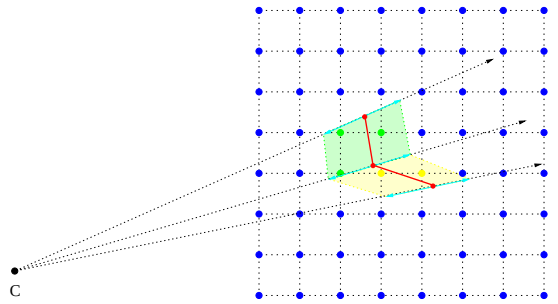

Figure 5: The two red line segments are the 2D equivalents to a patch of two triangles in 3D. Casting rays from the camera center **C** through the limiting points of one line segment, voxels in the area between these two rays need to be tested. Because the signed distance function is defined to fall of at a certain distance in front and behind the surface, the area can be further constrained to the light green area for the first line segment and the yellow area for the second.

to volumetric functions, which can be more easily and robustly fused. From the fused volume function a final triangle mesh is then retrieved by means of isosurface extraction. Our approach is based on an existing approach (Curless and Levoy, 1996) but improves the way partial reconstructions are converted to the volumetric function representation.

The volumetric function represents a surface by mapping points in R_3 to a vector in R_2, representing the signed distance of the point to the surface along the line of sight of the camera. Like Curless and Levoy we use a discrete representation of this volumetric function called a weighted signed distance (WSD) grid, which is a regular volume of samples of the continuous function. In theory the volumetric function should extend indefinitely from the surface in both directions. To avoid surfaces on opposite sides of an object to interfere with each other, it is forced to fall-off in a certain distance of the surface. To convert a triangle mesh to a WSD representation, for each WSD grid voxel a sampling value of the WSD function needs to be found. Curless and Levox propose a scan-conversion process for the conversion. We have, however, developed a both simple and efficient alternative to their approach: The WSD value for a voxel can be found by casting a ray from the camera center through the voxel and then intersecting it with the triangle mesh. Our approach exploits the fact that the WSD functions are only defined in a certain distance of the surface. Hence we obtain the WSD values only for voxels that are close enough to the surface. Figure 5 explains this concept.

Partial reconstruction in WSD representation can be merged easily and robustly using a additive scheme proposed by Curless and Levoy. The cumulative functions $D(x,y,z)$ and $W(x,y,z)$ of the merged WSD volume of n partial volumes are defined as

$$D(x,y,z) = \frac{\sum_{i=1}^{n} w_i(x,y,z) d_i(x,y,z)}{\sum_{i=1}^{n} w_i(x,y,z)},$$

$$W(x,y,z) = \sum_{i=1}^{n} w_i(x,y,z).$$

Rewritten as an incremental calculation, the cumulative signed distance and weight functions after merging the ith partial reconstruction read

$$D_{i+1}(x,y,z) = \frac{D_i(x,y,z) W_i(x,y,z)}{W_i(x,y,z) + w_{i+1}(x,y,z)} + \frac{D w_{i+1}(x,y,z) d_{i+1}(x,y,z)}{W_i(x,y,z) + w_{i+1}(x,y,z)},$$

$$W_{i+1}(x,y,z) = W_i(x,y,z) + w_{i+1}(x,y,z).$$

Merging auxiliary WSD volumes into the global WSD volume therefore is simply a matter of applying the rules defined above to each voxel.

A triangle mesh representation of the merged WSD volume can be obtained at any time during the sequential fusion of partial reconstructions from the WSD volume using the marching cubes isosurface extraction algorithm (Lorensen and Cline, 1987). The isosurface is defined as $D(x,y,z) = 0$. Therefore, voxel configurations can directly be found by testing the signs of the signed distance function $D(x,y,z)$ at the eight corners of a voxel. To skip empty space, only voxels which have nonzero weights $W(x,y,z)$ for all eight corners are processed.

3.4 Appearance Recovery

Recovering the shape of an object is not enough for a visually convincing reconstruction. Besides the shape, appearance information, like color and texture, is very important. We recover appearance by means of a simple texture mapping approach. We generate both texture coordinates and the texture map triangle by triangle: for each triangle all views in which its projection is front facing and not occluded are found. The triangle is then assigned to the view which ensures the highest possible texture resolution. Once all triangles are assigned to views, rectangular texture snippets corresponding to the bounding boxes of the projections of the triangles are obtained from the views and added to a list sorted by the width of the snippets. A texture map can now be populated with the texture snippets for each triangle, by traversing

Figure 6: Cutout of a texture image populated column by column with rectangular texture snippets for each shape triangle. The full texture has a resolution of 2048×2048 pixels.

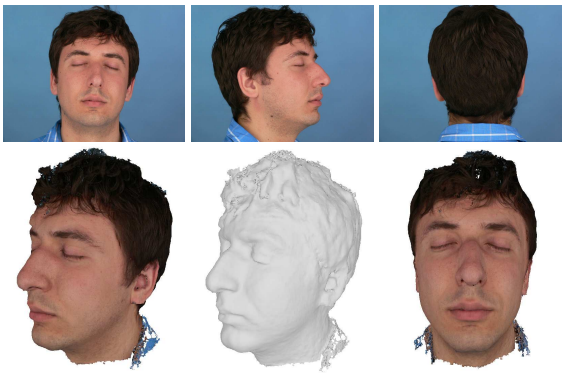

Figure 7: Sample images of a 90 view image sequence of a human head (top) and textured as well as untextured images of the reconstruction results (bottom).

the sorted list and filling the image with rectangular texture snippets column by column. Figure 3.4 shows a cutout of a generated texture image. Finally, the positions of the texture snippets are normalized to $[0, 1]$ to match texture mapping standards and stored as texture coordinates, defining the mapping.

4 EXPERIMENTS AND RESULTS

We have acquired and reconstructed more than 30 image sequences. The full reconstruction of a 90 view image sequence takes about two hours on a Pentium IV machine. A selection of typical results is shown in Figure 7, Figure 8 and Figure 9. While the reconstruction results are visually convincing we are also interested in an objective metric for the reconstruction accuracy. This however requires ground truth reference data (for example acquired using a laser scanner) which was not available to us. Hence we have at least tried to determine the robustness of our approach. For this purpose. we have acquired and reconstructed three different image sequences of the same person. The reconstruction results should be and as Figure 10 shows are very similar.

Figure 8: Sample images of a 85 view image sequence of a piece of wood (top) and textured as well as untextured images of the reconstruction results (bottom).

Figure 9: Sample images of a 92 view image sequence of a plastic dinosaur (top) and textured as well as untextured images of the reconstruction results (bottom).

5 CONCLUSION

We have developed a passive 3d reconstruction approach based on structure from motion and shape from stereo concepts. While the results are visually convincing we have not yet verified the reconstruc-

Figure 10: Reconstructions of three different image sequences of the same object show the robustness of our approach.

tion accuracy using ground truth data and an objective metric. Furthermore our reconstruction approach suffers from several principle shape from stereo problems like untextured areas, occlusions and repetitive patterns. Hence it can not be used for the reconstruction of untextured or specular objects. To solve these problems we are working on acquiring ground truth data using artificial image sequences generated from textured 3d models and we try to combine our shape from stereo based approach with shape from silhouette concepts to improve the reconstruction quality of sparsely-textured objects.

REFERENCES

Andrew W. Fitzgibbon, Geoff Cross, A. Z. (1998). Automatic 3d model construction for turn-table sequences. In *3D Structure from Multiple Images of Large-Scale Environments: European Workshop*, page 155.

Brown, D. (1976). The bundle adjustment - progress and prospect. In *XIII Congress of the ISPRS*, Helsinki.

Brown, M., Burschka, D., and Hager, G. (2003). Advances in computational stereo. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 25, pages 993–1008.

Curless, B. and Levoy, M. (1996). A volumetric method for building complex models from range images. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312. ACM Press.

Favaro, P. and Soatto, S. (2002). Learning depth from defocus. In *ECCV 2002 : 7th European Conference on Computer Vision*, page 735ff, Copenhagen, Denmark.

Frischholz, R. W. (1997). *Beitrge zur automatischen dreidimensionalen Bewegungsanalyse*. PhD thesis, Universitt Erlangen Nürnberg.

Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Kutulakos, K. N. and Seitz, S. M. (1998). A theory of shape by space carving. Technical Report TR692.

Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(2):150–162.

Lorensen, W. E. and Cline, H. E. (1987). Marching cubes: A high resolution 3d surface construction algorithm. In *SIGGRAPH '87: Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 163–169. ACM Press.

Oram, D. (2001). Rectification for any epipolar geometry. In *British Machine Vision Conference (BMVC)*, pages 653–662.

Rodehorst, V. (2004). *Photogrammetrische 3D-Rekonstruktion im Nahbereich durch Auto-Kalibrierung mit projektiver Geometrie*. PhD thesis, TU Berlin.

Rupp, S., Elter, M., Breitung, M., Zink, W., and Küblbeck, C. (2006). Robust Camera Calibration using Discrete Optimization. *Enformatika Transactions on Engineering, Computing and Science*, 13:250 – 254.

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1/2/3):7–42.

Schechner, Y. Y. and Kiryati, N. (2000). Depth from defocus vs. stereo: How different really are they? *Int. J. Comput. Vision*, 39(2):141–162.

Shi, J. and Tomasi, C. (1994). Good features to track. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 593–600, Seattle.

Stefano, L. D. and Mattoccia, S. (2000). Fast stereo matching for the videt system using a general purpose processor with multimedia extensions. In *CAMP '00: Proceedings of the Fifth IEEE International Workshop on Computer Architectures for Machine Perception (CAMP'00)*, page 356. IEEE Computer Society.

Tarini, M., Callieri, M., Montani, C., Rocchini, C., Olsson, K., and Persson, T. (2002). Marching intersections: An efficient approach to shape-from-silhouette. In *VMV*, pages 283–290.

Tomasi, C. and Kanade, T. (1991). Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University.

Weik, S. (2000). A passive full body scanner using shape from silhouettes. In *International Conference on Pattern Recognition*, volume 1, pages 750 – 753.

Wolfgang Niem, J. W. (1997). Automatic reconstruction of 3d objects using a mobile monoscopic camera. In *International Conference on Recent Advances in 3-D Digital Imaging and Modeling*, pages 173 – 180.

Zabih, R. and Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *ECCV '94: Proceedings of the third European conference on Computer Vision*, volume 2, pages 151–158, Stockholm, Sweden. Springer-Verlag New York, Inc.

Zhang, R., Tsai, P.-S., Cryer, J. E., and Shah, M. (1999). Shape from shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706.

Zhang, Z. (1998). A flexible new technique for camera calibration. Technical Report MSR-TR-98-71, Microsoft Corporation.

Ziou, D. (1998). Passive depth from defocus using a spatial domain approach. In *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, page 799, Washington, DC, USA. IEEE Computer Society.