3D Modeling from Turntable Sequences Using Dense Stereo Carving and Multi-view Consistency

Jordi Arnabat
Hewlett Packard
Av. Graells 501, 08190
Sant Cugat, Catalonia
jordi.arnabat@hp.com

Selma Casanovas
Fico Mirrors, SA.
Can Magarola 08100
Mollet Vallès, Catalonia
selma.casanovas@ficosa.com

Gérard Medioni
IRIS
University of Southern California
90089-0273, Los Angeles, CA
medioni@usc.edu

Abstract

This paper addresses the problem of reconstructing a 3-D model, with texture, from a sequence of images of the object taken from a turntable. Previous approaches have successfully solved the problem either using active interaction techniques, like the laser pen in Immersion's LightScribe commercial system or directly relying on objects texture (not implemented commercially). In contrast, we present a general methodology, which combines silhouette carving with dense stereo and multi-view consistency. It functions even in the presence of highlights or lack of texture. Results are presented on real objects with the above-mentioned characteristics.

1. Introduction

As many applications that simulate interaction with real world [14] are demanding ever more realistic 3D models of objects, inferring 3D shape from multiple images of an object has become a key area of research in computer vision.

Among all 3D objects, of special interest are those considered geometrically complex, often presenting occlusions, concavities and non-smooth morphology. As pointed out by Kutulakos [10], significant progress was done on the topic of reconstructing an unknown scene from images under different situations: inaccurate calibration [6], non-stationary scenes [6] and coarse reconstruction when geometry is too complex [15, 18], but a general solution is still beyond the current state of the art.

As an additional contribution toward addressing this limitation, we propose a passive multi stereo technique to compute dense 3D information and resolve complex objects that usually exhibit significant occlusions and may contain both texture and textureless surface regions. Moreover, an automated method is presented to obtain a photo-realistic 3D model of an object by carving dense details into an initial coarse 3D model. We have implemented and tested this technique in a controlled environment where the camera calibration is known.

In this paper, first we present a general overview of the approach. Then, we discuss the key pieces of the algorithm, and finally we show some results obtained from real data.

2. Related Work

During the past, several methods have been proposed to construct 3D models of objects from images. As pointed out by Fitzgibbon [5], numerous techniques dealt with the construction of 3D models by silhouette intersection. A good example is the method proposed by Szeliski [17] that was commercially implemented in Geometrix's 3Scan technology. However, volume intersection techniques are limited by the visual hull effect, which does not guarantee and accurate reconstruction of the object due to concavities.

On the other hand, active techniques rely on the interaction of a known light source over the object to recover its structure: interferometry techniques, pattern light projection and weak-structured light projection[1] are some examples. All of them work well in controlled environments, achieving good results in complex shape recovery.

Also, a combination of both approaches, volume intersection and active laser carving, is implemented in Immersion's LightScribe commercial system, yielding 3D texture mapped models of impressive quality.

In the opposite direction, in 1998, Kutulakos [11] introduced a passive approach called space-carving that formulates shape recovery as a constraint satisfaction problem. Note that this technique does not need any external light source to obtain the fine details but requires a large number of images to achieve similar results.

In summary, while some techniques require an active interaction with external elements (e.g. laser, light patterns, weak structured light,...), others rely only on the original images but place constraints on the type of objects that can

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

be modeled.

3. Overview of the approach

In our experiment, we use a calibrated system composed by a video camera, a turn table and a back-light illumination system. The turn table allows us to take several views of the object by rotating it around a single axis at known positions. Once the system is set up for a given object, a calibration algorithm [20] is used to compute the relative motion between the image plane and the object reference frame (further referred to as world reference frame). Two images are taken for each given position of the turn table, one using back-light illumination and the other under normal light conditions. As a result, two collections of images, named back-light and textured, are the inputs for the algorithm.

First, an initial (coarse) 3D model from the intersection of the object silhouettes is computed. For each pair of views, the epipolar geometry, mathematically expressed by the Fundamental matrix, can be recovered from the calibration data [12]. The Fundamental matrix is then used by a passive dense-stereo matching algorithm to find correspondent pairs of points between textured pairs of images.

At this point, N disparity maps store the correspondent points in the original set of images. For each pair of matched points, we recover the correspondent 3D point by intersection of the backprojections. Doing the same for each disparity map, a set of dense 3D point-clouds can be obtained. By taking enough views, it is feasible to entirely cover the object surface. Also, redundancy from overlap between the views helps discard false stereo matches. We propose a method based on multiple view agreement to enforce consistency through all the views and keep only those 3D points likely to be on the object surface.

The resulting point-clouds are carved into the coarse 3D model [8], reconstructing a more precise representation of the object where the concavities and occlusions have been recovered. Finally, texture can be mapped from the original textured images to the model surface delivering photo-realistic appearence to the final model.

3.1. Coarse 3D model

The first step in our approach consists in building a coarse 3D model. Fitzgibbon [5] described a system which, given a sequence of images of an object rotating about a single axis, generates a 3D model automatically. The 3D model is computed as the intersection of the outline cones back-projected from all the views. 3D models resulting from silhouette intersection are limited by the visual hull effect, hence the surface of complex objects can be roughly recovered.

Mainly, there are two strategies to recover the surface details (concavities): active techniques, like the one proposed by Bouguet [1], or passive techniques, as the proposed by Kutulakos [11]. In this paper we propose to use passive stereo to obtain dense 3D information and refine the rough 3D model.

3.2. Dense Stereo

In a calibrated system, recovering the epipolar geometry is straightforward. It is out of the paper scope to describe the calibration process in detail, the reader is referred to [20, 4, 8] for detailed discussion. Our system is calibrated using a pin-hole camera model. Therefore, 13 parameters have been computed to recover the projective geometry: 5 (internal) parameters for the camera, 6 (external) parameters to reference the camera to the object frame reference and finally 2 parameters to model radial distortion of the lens. Since the system consists of one camera, different views of the object are obtained by rotating it around the rotation axis. This is is equivalent to a system where the object is fixed and the camera moves around it. In such a system, the epipolar geometry is recovered from the relative motion between two camera positions [12, 7].

A dense stereo algorithm is used to match correspondent points in a pair of images. Most dense stereo algorithms use the epipolar constraint to convert the original 2D matching problem into 1D, reducing its complexity. Still, a relaxation technique is needed to discard false positives. In our experiment, we have implemented an algorithm proposed by Quian Chen and Gérard Medioni [2]. The method is based on a multi-resolution approach allowing correlation along the epipolar lines to be computed only at the coarsest level. The disparity map is then propagated to the next level, where it is used as a starting point for correlation refinement along an interval of 5 pixels. Chen-Medioni's approach provides a fast and robust-to-noise algorithm to compute a dense disparity map.

Disparity maps are a concise representation of the correspondences between matched points in a pair of images. Recovering 3D information from them is a classical 2-view triangulation problem. Mathematically, triangulation is defined by a pair of projective equations for a pair of matched points. Using homogeneous coordinates, the problem is reduced to a linear system of equations [8]. Intersecting two lines in 3D space has no solution in general, an aproximation can be obtained by least squares minimization [8]. Pseudo-inverse and singular value decomposition are common methods to solve a linear system. Instead, we use Longuet-Higgins' formula [12], of special interest to reduce computational cost when intensive triangulation is required.

3.3. Multiple View Consistency

Correlation-based algorithms are known to be vulnerable to scenarios such as: lack of texture (which may be produced by either shadows and saturation), repetitive patterns and occlusions. These show up in the disparity map as ei-

ther holes (no match) or erroneous disparity values (mismatch). The latter yielding a wrong 3D point in space and producing artifacts in the reconstructed model. A method to discard false matches is needed in order not to carve the coarse model in excess. Existing methods such as tri-linear tensor [8] and Mellor [9] enforce consistency at the disparity level(2D); rather than Chou [3], who uses visibility constrains directly in the 3D space.

We propose a simpler approach, providing lower computational cost and reasonably good results: theoretically, a collection of 2D-matches of the same object point should reconstruct exactly the same 3D point. Due to accuracy limits in the camera calibration and the dense stereo algorithm, noise is introduced into the 3D reconstruction and the exact 3D point cannot be obtained. Note that mismatches will cause 3D errors larger than the inherent noise. Hence, a voxeling technique can be used as a discriminator between noise and mismatches, if a suitable voxel size can be defined. Given a camera, estimating the voxel size is straightforward when the size of the object is known, since the distance from the object to the camera is constrained by the need to maximize object resolution. Our approach works as follows: Consider the coarse 3D model divided in voxels and, for each view, compute the number of 3D points hosted in each voxel (impacts). Then, for every voxel, consider only the views with at least one impact. Compute the mean of impacts per view and discard views considered outliers. The visibility of each voxel is defined as the number of views that have not been discarded. Finally, 3D points contained in voxels with a visibility lower than two are not considered by the carving algorithm. Figure 1 shows an example where, due to a lack of texture in the original images, the agreement condition becomes essential in order not to destroy the silhouette model.

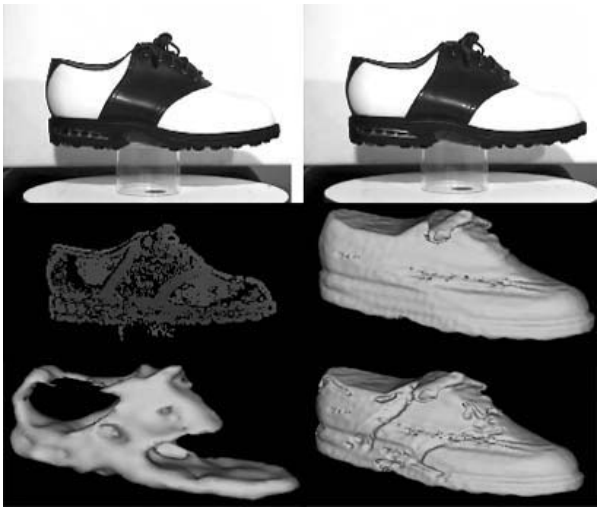

Figure 1. Original pair, disparity map, silhouettes based model, 36 views carved model and consensus carved model.

3.4. Carving and Texturing

Once the 3D information is reliable, a detailed 3D model can be obtained by carving the dense 3D points into the coarse 3D model. For each view, a 3D mesh is computed from the correspondent 3D point cloud. In our approach, a fast Delaunay-based triangulation [16] is used to save computational cost. The intersection volume from the silhouette-based model and the 3D meshes is eliminated from the coarse 3D model, resulting in a more accurate surface, which contains the fine details from the collection of views. Finally, texture is added to the carved model in order to provide a more realistic appereance. In our experiment, each triangle of the final mesh-model is projected back into the original set of images mapping the corresponding texture to the triangle face.

4. Results

The algorithm was tested within a variety of objects. We have found that 12 views are enough to ensure an accurate 3D models for simple objects (no occlusions). Between 12 and 36 views work well for the majority of complex objects, only in special cases, with both occlusions and lack of texture, taking more than 36 images has found useful. As a summary, figure 2 shows representative examples of the improvement achieved applying the technique. 36 views, a baseline of 10deg between the images, were used in all of them. Taking the entire process 8-10 minutes in an Intel P-III 800 Mhz. In our system the accuracy of the matching algorithm is around 0.7 pixels. Just as an example, in the dragon the estimated dimensions of the voxel were $1.4^3 mm^3$. The large concavity between the wings is recovered, as well as some fine details around the neck, head and legs. In this case, very few mismatches were obtained by the matching algorithm and hence, a small amount (2%) of 3D points were discarded by the agreement condition. The lighthouse also presented overall good texture except on some views where the tree hides a portion of the tower. On these views a shadow is projected over the tower reducing texture details. Hopefully, other views resolve the occlusion and the concavity got recovered as well. In the shoe example highlights and homogeneous-textured areas produce wrong matches yielding inconsistencies in 3D space. The agreement condition is only satisfied in certain regions, near the string and the seam, which are carved away. Without the agreement condition (see Figure 1), the original model would be completely destroyed. In this case, the amount of rejected 3D points reached 70%.

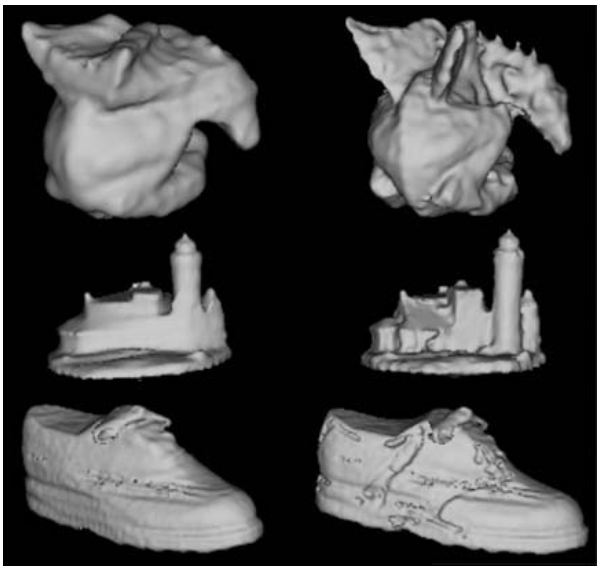

Figure 2. Silhouettes based (left) versus texture carved (right) models on different types of objects: strong texture (dragon, lighthouse), highlights and weak texture (shoe).

5. Conclusions

We have integrated three independent but complementary techniques, silhouette intersection, dense stereo matching and multiple view carving, provinding a fully automated method to build photo-realistic 3D model of objects.

Silhouette intersection provides a rough 3D model, but cannot resolve concavities. Dense stereo is used as a passive method to recover surface details. A multiple view agreement avoids non-consistent data being carved into the final model, enabling our technique to generate good models even in the presence of highlights and uniform texture.

Moreover, our method does not rely on active user intervention, resulting in a faster overall process compared to active techniques such as hand-held laser carving.

In this paper we have described the implementation of our approach in a calibrated environment. An interesting future direction is its extension to free-hand motion. Whilst most of the core algorithms used in our approach are directly extensible to non-calibrated environments, epipolar geometry will have to be recovered in absence of calibration data. This is proved feasible using robust feature matching [19, 13]. Therefore, future work should be centered in understanding whether the resulting overall process can produce reliable results.

6. Acknowledgment

This research was carried out while the authors were at Geometrix, Inc. The authors thank Geometrix colleagues for their insights, support and collaboration in many related projects over 2000-2001. Special thanks go to Mike Tsoupko-Sitnikov, Franco Callari, Oleg Mishin, Jin-Long Chen, David Guigonis, Bastien Pesenti, Roman Waupotitsch and Arthur Zwern.

References

[1] J. Bouguet and P. Perona. 3d photography on your desk. *ICCV*, pages 43–49, 1998.

[2] G. Chen and G. Medioni. A volumetric stereo matching method: Application to image-based modeling. *CVPR*, I:29–34, 1999.

[3] G. Chou and S. Teller. Multi-level 3d reconstruction with visibility constraints. *DARPA*, pages 543–550, 1998.

[4] O. Faugeras. *Three-dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, 1993.

[5] A. Fitzgibbon, G. Cross, and A. Zisserman. Automatic 3d model construction for turn-table sequences. *Lecture Notes in Computer Vision Science*, (1506):155–170, 1998.

[6] P. Fua and Y. Leclerc. Registration without correspondences. *CVPR*, pages 121–128, 1994.

[7] R. Hartley. In defense of the eight-point algorithm. *PAMI*, 19(6):580–592, June 1997.

[8] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, June 2000.

[9] M. J.P., S. Teller, and T. Lozano-Perez. Dense depth maps from epipolar images. *MIT AI Memo*, 1593:893–900, 1995.

[10] K. Kutulakos. Approximate n-view stereo. *ECCV*, 1:67–83, July 2000.

[11] K. Kutulakos and S. Seitz. A theory of shape by space carving. *IJCV*, 38(3):199–218, July 2000.

[12] H. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, September 1981.

[13] Q. Luong and F. O.D. The fundamental matrix: Theory, algorithms and stability analysis. *IJCV*, 17(1):43–75, January 1996.

[14] P. Narayanan, P. Rander, and T. Kanade. Constructing virtual worlds using dense stereo. *ICCV*, pages 3–10, 1998.

[15] A. Prock and C. Dyer. Towards real-time voxel coloring. *Image Understanding Workshop*, pages 315–321, 1998.

[16] J. Shewchuck. *Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator*. Springer-Verlag, May 1996.

[17] R. Szeliski. Shape from rotation. *CVPR*, pages 625–630, 1991.

[18] R. Szeliski. Rapid octree construction from image sequences. *CVGIP*, 58(1):23–32, July 1993.

[19] Z. Zhang. Determining the epipolar geometry and its uncertainty: A review. *IJCV*, 27(2):161–195, 1998.

[20] Z. Zhang. A flexible new technique for camera calibration. *PAMI*, 22(11):1330–1334, November 2000.