3D Scene Reconstruction from Multiple Uncalibrated Views

Li Tao
ltao2@stanford.edu

Xuerong Xiao
xuerong@stanford.edu

Abstract

In this project, we focus on the problem of 3D scene reconstruction from multiple uncalibrated views. We have studied different 3D scene reconstruction methods, including Structure from Motion (SFM) and volumetric stereo (space carving and voxel coloring). Here we report the results of applying these methods to different scenes, ranging from simple geometric structures to complicated buildings, and will compare the performances of different methods.

1. Introduction

3D reconstruction from multiple images is the creation of three-dimensional models from a set of images. It is the reverse process of obtaining 2D images from 3D scenes. In recent decades, there is an important demand for 3D content for computer graphics, virtual reality and communication, triggering a change in emphasis for the requirements. Many existing systems for constructing 3D models are built around specialized hardware (e.g. stereo rigs) resulting in a high cost, which cannot satisfy the requirement of its new applications. This gap stimulates the use of digital imaging facilities (such as cameras) [1].

The 3D scene reconstruction from multiple view images is an increasingly popular topic which can be applied to street view mapping, building construction, gaming and even tourism etc. When the reconstruction of a 3D scene is needed, a reliable computer vision based reconstruction method is much more cost-efficient and time-efficient than traditional methods such as aerial photo filming. The 3D scene reconstruction applications such as Google Earth allow people to take flight over entire metropolitan areas in a virtually real 3D world, explore 3D tours of buildings, cities and famous landmarks, as well as take a virtual walk around natural and cultural landmarks without having to be physically there. A computer vision based reconstruction method also allows the use of rich image resources from the internet.

In this project, we have studied different 3D scene reconstruction methods, including Structure from Motion (SFM) method and volumetric stereo (space carving and voxel coloring). Here we report the results of applying these methods to different scenes, ranging from simple geometric structures to complicated buildings, and will compare the performances of different methods.

2. Previous Work

2.1. Review of Previous Work

Reconstructing 3D models just by taking and using 2D images of objects has been a popular research topic in computer vision due to its potential to create an accurate model of the 3D world with very low cost. 3D reconstruction from one image alone is possible [2,3] but performs not as good as reconstruction from multiple images with different views. In most practical cases, the images used to reconstruct a 3D model are not calibrated. Therefore structure from motion becomes a very popular and convenient method to calculate both camera parameters and 3D point positions simultaneously, only depending on the identified correspondences between the

2D images used. Many successful applications have been created using this concept, such as the Photo Tourism project, which establishes a system for interactively browsing and exploring large unstructured collections of photographs of a scene using a novel 3D interface and successfully reconstructs various historic and natural scenes [4].

Another genre of 3D reconstruction methods is volumetric stereo, including space carving [5,6] and voxel coloring [7]. Unlike SFM, these methods do not require the identification of a large number of correspondences for dense 3D reconstruction. However, they require calibrated images, i.e., both the intrinsic and extrinsic parameters of the cameras associated with every view must be known and input to the algorithm for successful reconstruction.

2.2. Our Work

In this project, we have studied several 3D reconstruction methods, including both SFM and volumetric stereo. We have both implemented our own algorithms and tried out several software packages and applications available. We have tested their performances for reconstructing different scenes ranging from small objects to large buildings. Here we will report the results and compare the performance of different methods.

3. Technical Part

3.1. Summary of Technical Solutions

In this project, we have implemented both SFM and volumetric stereo algorithms for the 3D reconstruction task. After camera calibration, we achieved SFM from two views [8] and multiple views [9]. We also experimented with the Princeton SFMedu package [10] extensively using images of multi-scale objects. The package contains an additional multi-view stereo step and produces dense reconstruction from the output of the SFM algorithm. In volumetric stereo, we focus on space carving for reconstructing 3D visual hulls. In order to acquire calibrated views, we obtained camera parameters (intrinsic and extrinsic) from our SFM algorithm ad used these parameters as the input for the space carving algorithm. We have also tried out applications employing voxel coloring [11] and a smart phone based commercial 3D reconstruction application [12] for comparison. The details are discussed below.

3.2. Technical Solutions

3.2.1 SFM from two views

We first implemented the SFM algorithm in Matlab and started with two views [8]. Using the Camera Calibrator app, we calculate the camera intrinsic parameters from 16 pictures of an asymmetric checkerboard pattern. The 2D and 3D information are linked by measuring the grid size of the checkerboard. Figure 1(a) below shows the automatic checkerboard corner detection. Figure 1(b) shows the reprojection errors and Figure 1(c) shows the picture orientations (thus camera positions) in the world coordinate system. We eliminate some pictures by setting a threshold on the mean reprojection error. The resulting camera intrinsics, the lens distortion coefficients and other parameters are then input to the SFM process.

Then we used the calibrated camera to take two pictures of the same object in two different views. We extract features and find point correspondences between the two images. We have experimented with Harris features, FAST features, and SURF features etc. The SURF features give rise to the best result with the target scenes. With the identified correspondences, we then estimate the fundamental matrix and find the epipolar inliers using RANSAC method. We calculate the camera positions using these inliers. With the calculated camera positions, we perform the dense reconstruction of the 3D scene using triangulation as the last step.

3.2.2 SFM from multiple views

Using the same calibrated camera, we moved from two-view reconstruction to reconstruction with multiple views. We need to merge the point clouds during the reconstruction process. Bundle adjustment is hence performed to refine the resulting co-

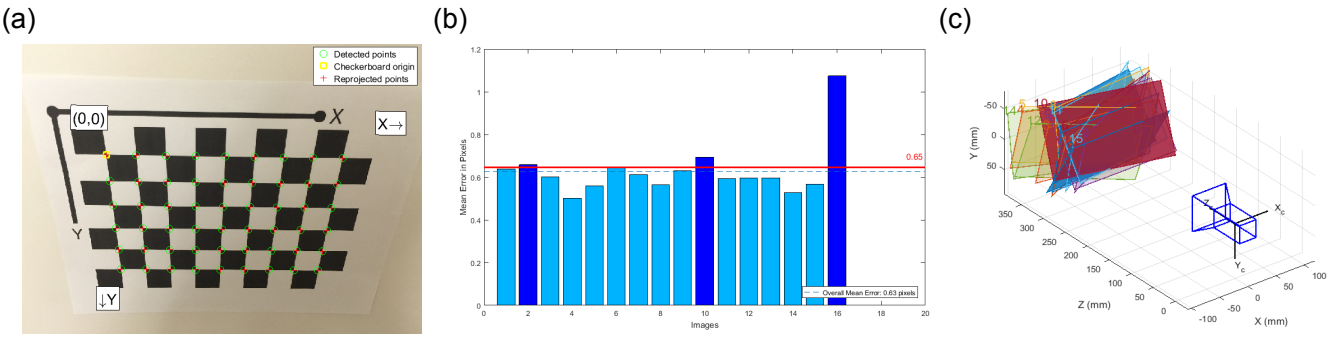

Figure 1. (a) Automatic checkerboard corner detection; (b) Reprojection errors of different pictures; (c) Picture orientations (camera positions) in the world coordinate system.

ordinates by minimizing the reprojection error [9].

3.2.3 Princeton SFMedu

There are four steps of image-based modeling: SFM, multi-view stereo, model fitting and texture mapping [10], following a bottom-up order. From images taken from moving cameras, the SFM method computes 3D point cloud, after which the multi-view stereo generates denser points. Model fitting creates meshes from points, and texture mapping leads to models based on the meshes. Up to this point, we reconstruct scenes using SFM and multi-view stereo.

Given two images taken by a moving camera at different locations, keypoints are described using SIFT features. The keypoints in the two images are then matched to estimate the fundamental matrix and essential matrix. RANSAC is used to eliminate outliers and prune the estimation. The camera trajectory described by the rotation and translation matrix is then computed. For multi-view reconstruction, bundle adjustment is performed to refine the resulting coordinates by minimizing the reprojection error. Multiple view stereo is implemented using matching propagation, which utilizes the zero-mean normalized cross-correlation in a priority queue and add new potential matches in each iteration. The algorithm propagates only on areas with maximal gradient greater than a threshold. Lastly, the triangulation step is used to generate colorized 3D point clouds.

3.2.4 Space carving

In order to acquire calibrated views for space carving, we first used the SFM algorithm to calculate the intrinsic and extrinsic camera parameters associated with different views, and then used these parameters as the input to the space carving algorithm. In order to acquire better silhouettes, we have manually removed the background of input images. The space carving algorithm is developed based on the code from course homework.

3.2.5 Voxel coloring

We have also experimented with a generalized voxel coloring implementation - the Archimedes executable [11]. This implementation also requires calibrated views (preferably images acquired from an object placed on a turntable) and accurate intrinsic and extrinsic camera parameters. We have tried out a provided example with this implementation and will discuss its performance in later sections.

3.2.6 Commercial 3D reconstruction application - Autodesk 123D Catch

For comparison purposes, we also experimented with a commercial smart phone based 3D reconstruction application - Autodesk 123D Catch [12]. This application does not require calibrated views. A dense and accurate 3D model can be generated just with images taken from different views of an object.

4. Experiments and Results

4.1. SFM from Two Views

With SFM from two views, the identified SURF features of the first image are shown in Figure 2(a). The tracked correspondences between two images are shown in Figure 2(b). The resulting identified epipolar inliers are shown in Figure 2(c). And the dense reconstruction of the 3D scene along with the camera positions are shown in Figure 2(d).

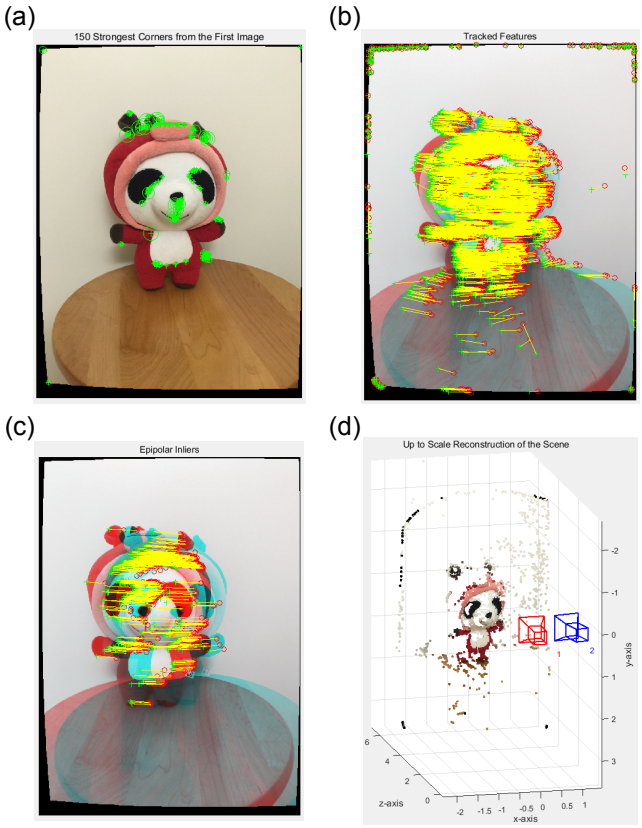

Figure 2. (a) 150 strongest SURF corners from the first image; (b) Tracked correspondences between two images; (c) Identified epipolar inliers; (d) Dense reconstruction of the 3D scene and the camera positions.

4.2. SFM from Multiple Views

With SFM from multiple views, we can also acquire the dense 3D reconstruction of the same scene as used in two-view SFM. The original input image sequence is shown in Figure 3(a) and the dense 3D reconstruction of the scene (from two viewing angles) is shown in Figure 3(b).

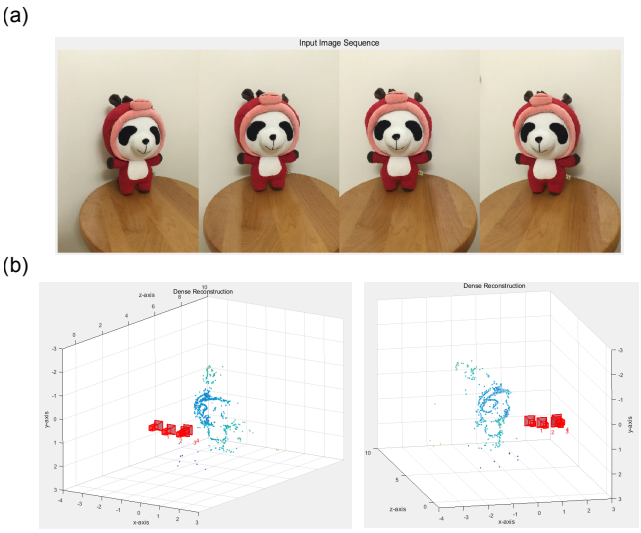

Figure 3. (a) Original input image sequence; (b) Dense 3D reconstruction of the scene (from two viewing angles).

By adjusting the threshold of the number of key features to keep, the density of the reconstructed points varies. When the threshold is low, the reconstruction is dense with more outliers; when the threshold is high, the reconstruction is more sparse with fewer outliers. In multi-view SFM, it is more difficult to extract correspondence epipolar inliers across all the views, thereby resulting in less dense reconstruction compared to the two-view SFM case. However, the advantage of multi-view SFM is that since we input images from more viewing angles, the reconstructed model reveals more 3D information about the object.

4.3. Princeton SFMedu

The Princeton SFMedu package also features SFM reconstruction with multiple uncalibrated views. The reconstruction results of various scenes are shown in Figure 4, with the left part showing the original input image sequence and the right part showing the reconstructed 3D models. The difference between this package and our SFM algorithm is that it uses SIFT features instead of SURF features. It also creates a dense 3D reconstruction as a "patchwork" of the input image segments, instead of a collection of bare 3D points,

which makes the reconstructed model looks more like the real object.

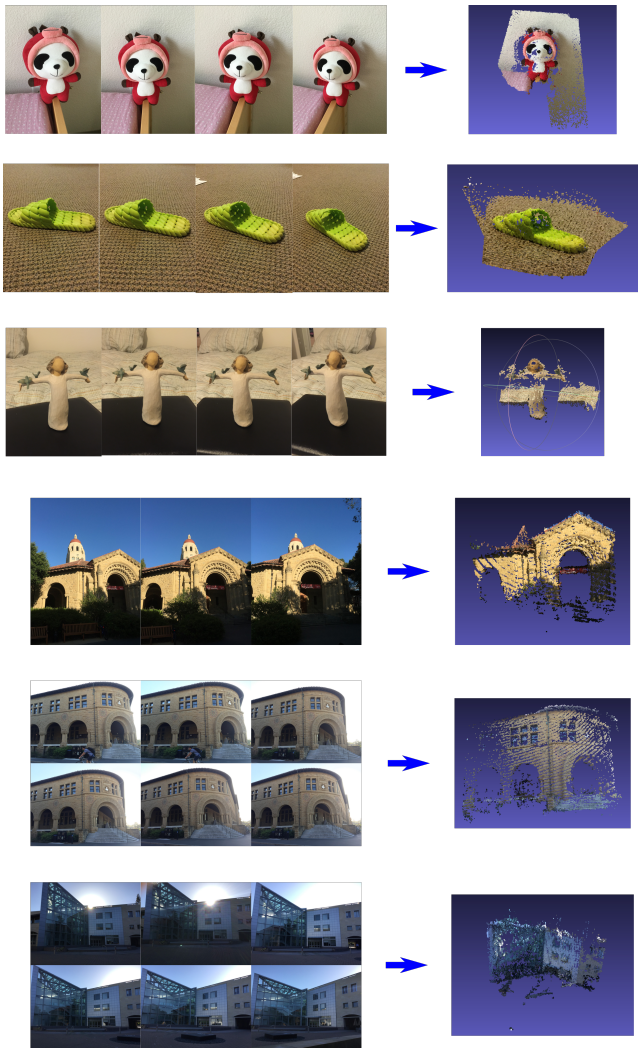

Figure 4. Reconstruction results of various scenes with Princeton SFMedu package. The 3D scene snapshots are taken from MeshLab.

4.4. Space Carving

The intrinsic and extrinsic camera parameters required for space carving are acquired with our multi-view SFM algorithm. We have also manually removed the background of input images for better silhouette generation. One of the original input images (after background removal) and the associated silhouette with it are shown in Figure 5. The reconstruction result after one carving is shown in Figure 6(a). The result after all (four) carvings is shown in Figure 6(b). The final carving result with input camera positions and associated images is shown in Figure 6(c).

Figure 5. One of the original input images for space carving and its associated silhouette.

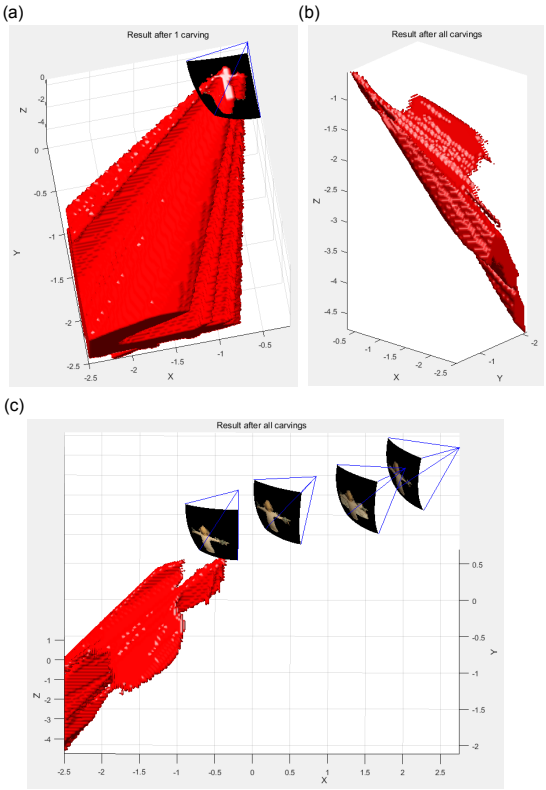

Figure 6. (a) Reconstruction result after one carving; (b) Reconstruction result after all carvings; (c) Final carving result with input camera positions and associated images.

In the results shown above, we have enlarged the input images in order to clearly show the projection of the visual hull onto the images. From

Figure 6(a), we can see that the ”cross” structure of the object is clearly shown after one carving. However, after all four carvings, the reconstructed model does not accurately represent the original object. This is because the camera parameters acquired with the SFM algorithm are not accurate. From Figure 6(c), we can see that these acquired camera positions are not consistent with each other, resulting in the cut-out of some parts of the visual hull that actually belong to the object. We have found that slight inaccuracy in the camera parameters can lead to large error in the reconstructed model with space carving.

4.5. Voxel Coloring

The reconstruction result of a provided example (a sitting teddy bear) in the Archimedes executable [11] which employs the generalized voxel coloring concept is shown in Figure 7. The result is shown from two viewing angles, both in cubes (Figure 7(a)) and with smooth surface (Figure 7(b)). Since this implementation requires a specific input file type, we did not have the time to experiment with our own images. This will be a future step of our work.

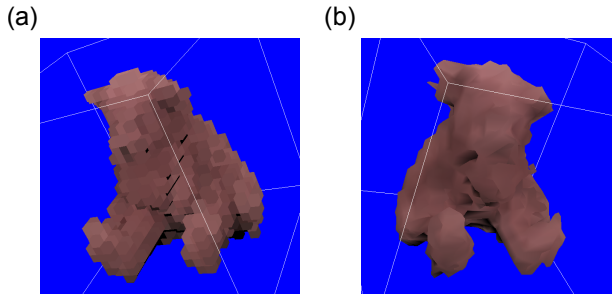

Figure 7. Reconstruction result of a sitting teddy bear in the Archimedes executable [11] shown (a) in cubes and (b) with smooth surface.

4.6. Commercial 3D Reconstruction Application - Autodesk 123D Catch

As the last step, we experimented with the smart phone based 3D reconstruction application - Autodesk 123D Catch [12]. Two reconstruction examples are shown in Figure 8. The left part shows one of the input images (front view) and the right part shows the reconstruction result from a few different viewing angles. Each reconstruction requires around 20-30 input images taken from different views.

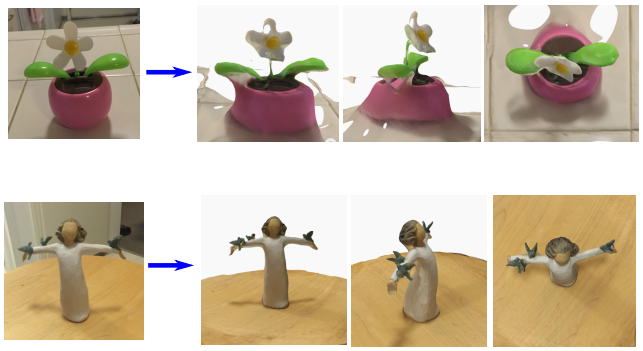

Figure 8. Reconstruction examples with Autodesk 123D Catch [12].

Compared to other methods studied, the Autodesk 123D Catch application gives the best and most robust reconstruction of 3D scenes. For example, we have tried to use the SFM algorithm to reconstruct the first scene (flower). However due to the small number of identified inlier correspondences, the reconstruction was not successful. The Autodesk 123D Catch application also makes use of image refining methods to make the reconstructed model look very similar to the real object. The disadvantage of this application is that it tends to take a long time (around one hour) for 3D scene reconstruction with 20-30 input images.

5. Conclusions

To achieve 3D reconstruction from uncalibrated views, we apply SFM, space carving, and voxel coloring to different scenes with a wide range of geometric complexity. We also implement several off-the-shelf packages and examine the performance in different settings.

In the two-view SFM, we compute the camera parameters by finding point correspondences and estimating the fundamental matrix. In the multi-view SFM, we merge and refine the point clouds using bundle adjustment. The reconstructed result is denser with two-view SFM due to the fact that it is difficult to find correspondence inliers across all

the views in multi-view SFM. However, the resulting model from the multi-view reconstruction possesses more 3D spatial information. In comparison, with an additional multi-view stereo step that iteratively searches for nearby patches, the Princeton SFMedu package can reconstruct models that resemble the real target objects.

Our space carving algorithm takes the camera parameters from our SFM algorithm as input. The one-carving result reveals the basic structure of the 3D object. The multiple-carving result is too susceptible to the inaccuracy of the input camera parameters to preserve the entire 3D structure. We also test a voxel coloring package with a provided example and will further the investigation with our own images in the future.

Despite the long running time, the commercial Autodesk 123D Catch application provides robust and accurate 3D models. This is partly due to its ability of integrating a relatively large number of images and its refining algorithm for the reconstructed model.

Please find our project code in this link: **https://drive.google.com/open?id=0B0A7ShsB5wBgYTR0QVlwVTdPRVk**.

6. References

[1] https://en.wikipedia.org/wiki/3D_reconstruction_from_multiple_images

[2] Saxena, Ashutosh, Sung H. Chung, and Andrew Y. Ng. "3-d depth reconstruction from a single still image." *International journal of computer vision* 76.1 (2008): 53-69.

[3] Saxena, Ashutosh, Min Sun, and Andrew Y. Ng. "Make3d: Learning 3d scene structure from a single still image." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31.5 (2009): 824-840.

[4] Snavely, Noah, Steven M. Seitz, and Richard Szeliski. "Photo tourism: exploring photo collections in 3D." *ACM transactions on graphics (TOG)*. Vol. 25. No. 3. ACM, 2006.

[5] Kutulakos, Kiriakos N., and Steven M. Seitz. "A theory of shape by space carving." *International Journal of Computer Vision* 38.3 (2000): 199-218.

[6] Fitzgibbon, Andrew W., Geoff Cross, and Andrew Zisserman. "Automatic 3D model construction for turn-table sequences." *3D Structure from Multiple Images of Large-Scale Environments*. Springer Berlin Heidelberg, 1998. 155-170.

[7] Seitz, Steven M., and Charles R. Dyer. "Photorealistic scene reconstruction by voxel coloring." Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on. IEEE, 1997.

[8] http://www.mathworks.com/help/vision/examples/structure-from-motion-from-two-views.html

[9] http://www.mathworks.com/help/vision/examples/structure-from-motion-from-multiple-views.html

[10] http://robots.princeton.edu/courses/SFMedu

[11] http://matt.loper.org/Archimedes/Archimedes_docs/html/index.html

[12] http://www.123dapp.com/catch