Understanding Geometry of Objects: Shape from Motion and Shading

by
Chahat Deep Singh
Robotics Graduate Student

***Abstract*—This project aim towards getting the shape, illumination, reflectance and other parameters from an image or a stream of images using Tomasi-Kanade's and Barron-Malik's method. Hence comparing the results between the two. Factorization approach did not show better results, especially on the Castle data set and thus Bundle Adjustment, state-of-the-art non linear approach was used for getting a structure from the stream of images. For getting the shape, reflectance, or illumination from shading there were certain constrains/assumptions: surfaces tend to be smooth, paint tends to be uniform and illumination tends to be natural. With these constrains, the shape of the object is constructed from a single image.**

I. Introduction

The objective of this project is to compare the results between the two stated algorithms. In order to do so, first the factorization approach was used to get the Structure from motion. Clearly, the results were not good enough. I could clearly see an erroneous structure in the result from Tomasi-Kanade's [1] approach. Though, for a technique published in 1992, it gives amazing results. This is because it assumes it to be a linear problem and does a simple SVD after normalizing. In real world scenario, almost everything is completely non-linear. Thus, I tried getting the structure from an image stream using Agarwal *et al.* [2] technique- Bundle Adjustment, a non-linear approach.

For Shape, Illumination and Reflectance from Shading; a 3D structure of the object is constructed from a single image. Recovering such properties from a single image seems almost impossible till 2015. Before 2015, a variety of findings were made for getting the shape from multiple images. It takes a single(masked) image of an object and produces the model of the same as output with a reasonable estimate of the shape, surface normals, reflectance, shading and illumination which produced that image. Barron-Malik's [3] method outperforms all previous solutions to such problems.

II. Shape and Motion from Image Stream: Factorization Approach & Bundle Adjustment

In this section, different kinds of algorithms to get 3D structure from multiple images. These algorithms were implemented on MATLAB. First, let us focus on the orthographic factorization approach of Tomasi-Kanade.

The algorithm says that given an image stream, let us say that we have tracked (using KLT) P feature points over F frames, we write the *measurement matrix* of size $2F$xP:

$$W = \begin{bmatrix} U \\ \overline{V} \end{bmatrix}$$

Also, the shape matrix of size 3xP is defined as

$$S = [s_1 \cdot s_2 .... s_p]$$

The main objective of the factorization method is to compute the matrices R and S.

The entire algorithm can be explained in five steps:
Step 1: Compute the Singular-value decomposition.
Step 2: Define $\hat{R} = O_1^{'}(\Sigma)^{1/2}$ and $\hat{S} = O_2^{'}(\Sigma)^{1/2}$
Step 3: Impose Metric Constraints on matrix Q, 3x3 matrix (constraints with x,y and z direction) and use Newton's method to solve the system of linear equations.
Step 4: Compute the rotation matrix R and the shape matrix S as:

$$R = \hat{R}\ Q, \quad S = Q^{-1}\ \hat{S}$$

Step 5: Align the camera reference system with the world reference system by:

$$R = R\ R_0, \quad S = R_0^T\ S$$

The algorithm was ran on the *Medusa* and *Castle* image sequences. Since, KLT tracker was used to obtain the feature tracking points, it was not able to track the feature points well especially for the castle sequences (as the image sequence was not continuous) and thus a bad output was generated. Thus, I moved to the state-of-the-art algorithm developed by Agarwal *et al.* [2].

Building Rome in a Day:
This algorithm (even used by Microsoft in Photosynth application) can match and reconstruct 3D scenes from a collection of photographs irrespective of the viewing angle and illumination. The pipeline uses existing state of the art of large scale matching and SFM algorithms, including SIFT, vocabulary trees, Bundle Adjustment and few other known techniques. Unlike the factorization method, it does not linearize and compute SVD, instead it solves the non-linear system of equations. This algorithm is much slower than the factorization method but gives much better structures. Let us see the results:

Fig. 1. Castle Image form the dataset

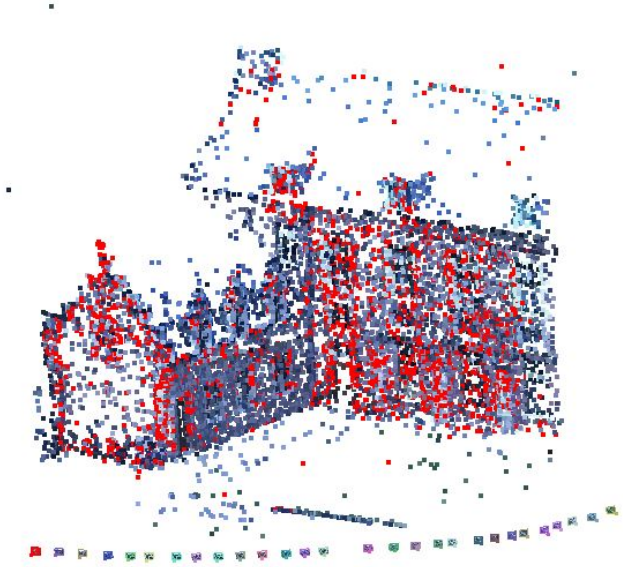

Fig. 2. Castle 3D Sparse Construction

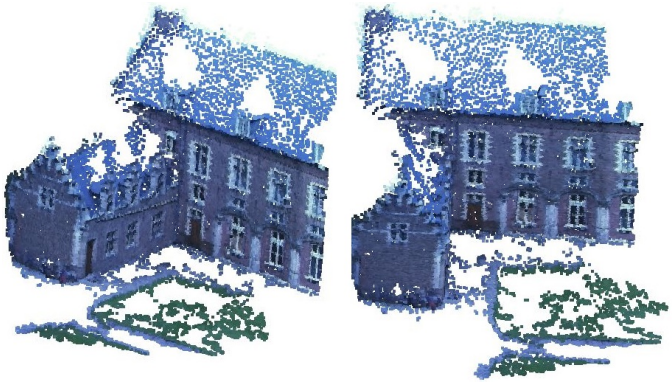

Fig. 3. Castle 3D Dense Construction

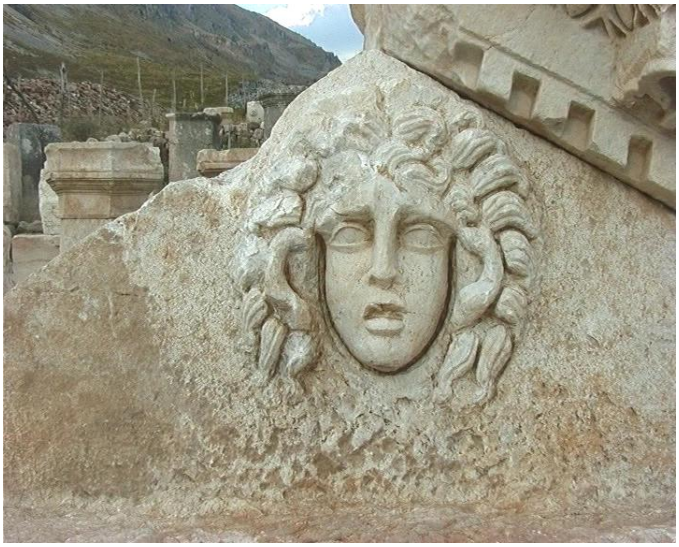

Fig. 4. Medusa Image form the dataset

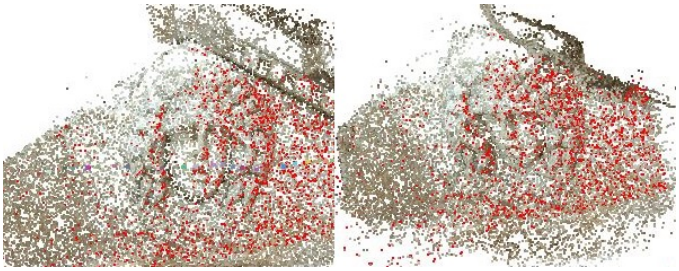

Fig. 5. Medusa 3D Sparse Construction

I also performed it on hand crafted data sets (Fig. 7,8,9). First, I tried on a 5 x 5 Professor's Rubik's Cube. With more than 400 images in the sequences, I was able to build an amazing dense representation of the dense structure.

Finally, I ran on our very own Testudo statue (from Stamp union, UMD). Even with a considerable amount of reflections on the statue, I was able to construct relatively good sparse and dense model of the statue. (Fig 10,11 and 12)

Fig. 6. Medusa 3D Dense Construction

Fig. 7. Rubik's Cube Image

Fig. 8. Rubik's Cube 3D Sparse Construction

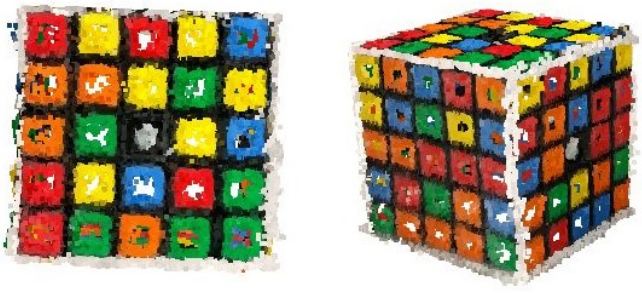

Fig. 9. Rubik's Cube 3D Dense Construction

Fig. 10. Testudo Image form Stamp Union

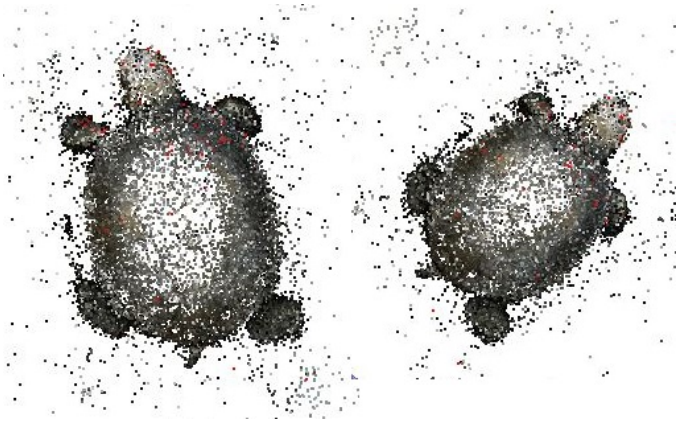

Fig. 11. Testudo 3D Sparse Construction

Fig. 12. Testudo 3D Dense Construction

III. Shape, Illumination and Reflectance From Shading

SIRFS is a model which takes a single masked input image of an object and produces a reasonable estimate of the shape, surface normals, reflectance, shading and illumination of the image. For a reasonably good estimate of the structural parameters, SIRFS has few assumptions: Surfaces tend to be isotropic and bend infrequency; Reflectance images tend to be piecewise smooth and low-entropy; Illumination tends to be natural. Assuming these priors and using Barron's multi-scale optimization technique, one can estimate the attributes of an object from a single image.

To estimate shape, illumination and reflectance, we must solve the optimization problem in equation:

$$minimize(Z, L): \quad g(I - S(Z, L)) \; + \; f(Z) \; + \; h(L)$$

where $g(R)$, $f(Z)$ and $h(L)$ are cost functions for reflectance, shape and illumination respectively. (Rest are standard variables from Barron's paper [3]. For this, Barron uses an effective multi-scale optimization technique, known as L-BFGS, which is simple to implement. Let us see the results:

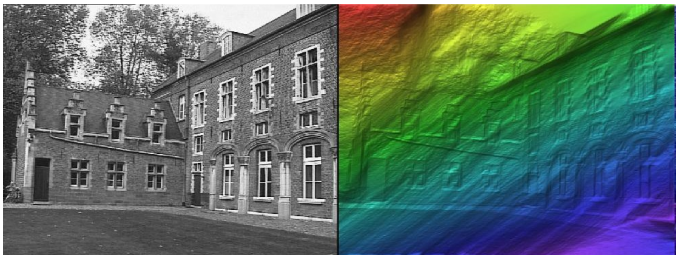

Fig. 13. Side View: Castle- Image and Shape

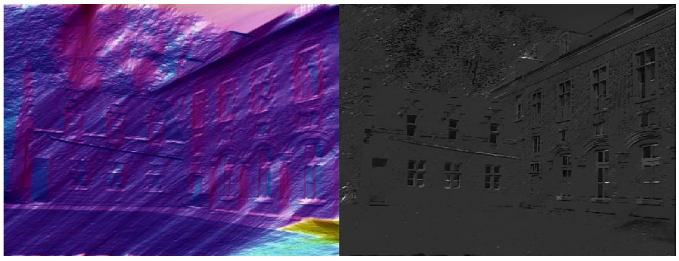

Fig. 14. Side View: Castle- Normals and Reflectance

Fig. 15. Side View: Castle- Shading and Light

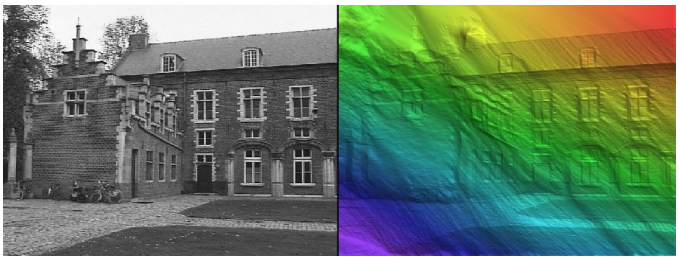

Fig. 16. Front View: Castle- Image and Shape

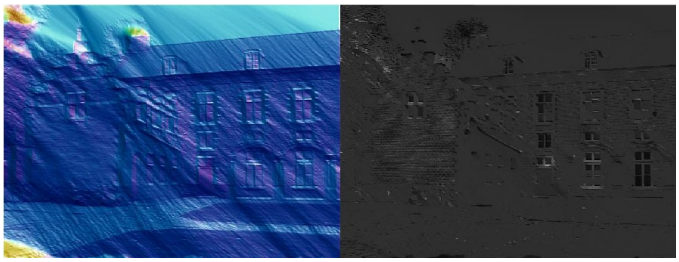

Fig. 17. Front View: Castle- Normals and Reflectance

Fig. 18. Front View: Castle- Shading and Light

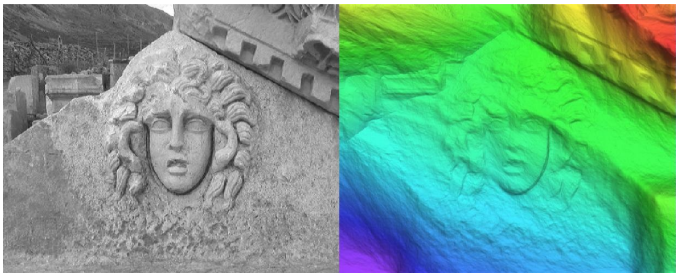

Fig. 19. Front View: Medusa- Image and Shape

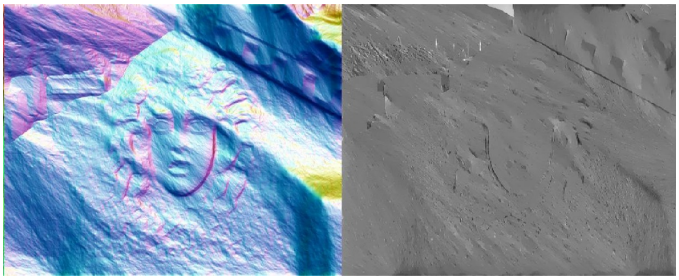

Fig. 20. Front View: Medusa- Normals and Reflectance

Fig. 21. Front View: Medusa- Shading and Light

Conclusion

Clearly, structure from motion fails for non-rigid objects. I tried implementing SFM on a swimming turtle. Only the 'head' and 'shell' of the turtle could be tracked (not the fins). It also fails when the KLT or any other tracker is not able to track the feature *i.e.* there should not be much difference between two consecutive frames. Also, to track features between two frames, the brightness of each feature should remain the same in consecutive frames. In the case of 'Shape from Shading', Barron mentions few priors which fails miserably on real world images. Definitely, Barron-Malik's method works amazingly well on their datasets but not on the other images. For every image, mask needs to be change and thus the algorithm is not robust enough unlike the SFM algorithm mentioned above. Also, since shading is an ingerently poor cue for low-frequency shape estimation, Barron-Malik' model often makes mistakes in coarse shape estimation, especially if there are multiple light sources very closeby to the object. Failure modes of both algorithms; reasons for superlative/poor performance.

References

[1] C Tomasi, T Kanade, "Shape and motion from image streams under orthography: a factorization method" International Journal of Computer Vision, 1992 - Springer

[2] S. Agarwal, N. Snavely, I. Simon S. Seitz and R. Szeliski "Building Rome in a day", Computer Vision, 2009 IEEE 12th International Conference.

[3] J. Barron, J. Malik "Shape, Illumination, and Reflectance from Shading", IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2015

[4] Changchang Wu, S. Agarwal, B. Curless, S. Seitz, "Multicore Bundle Adjustment", CVPR 2011

[5] E. Zheng, Changchang Wu, "Structure from Motion Using Structure-less Resection", ICCV 2015.