Sparse Three Dimensional Reconstruction using Structure from Motion

Ameya Shringi[1] Vsihal Garg[2]
[1,2]Golisano School of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York-14623

***Abstract*— The objective of the project is to generate a sparse three dimensional model using multiple images of an object. To achieve the same, structure from motion is implemented which relies on numerous image to incrementally compose a model based on the feature similarities in different images. These feature similarities are obtained using SURF and further pruned by adding camera constraint in form of computing fundamental matrix. The matrix obtained, along with the camera parameters are used to calculate essential matrix that enforces additional constraints on the matching points. The essential matrix is decomposed to determine motion between two images. Once motion coefficients are determined they are used to triangulate matching points in the third dimension.**

I. INTRODUCTION

Three Dimensional representation can be more expressive when looking at a large collection of unstructured photographs when compared with a linear arrangement. In 3D model reconstruction the aim is to create a 3 dimensional model of a place, object or any entity using various 2 dimensional images of the same.

The motivation behind this research is that if it is possible to create 2d photos of the 3-d world as we see it, can this process be reversed and 3d models can be created from 2d photos. It turns out that, given multiple images of an object taken from different angles, it is possible to create a 3d model of the object.

Apart from just being attractive, this technique can be very useful across various fields. For example: given multiple images of a historical place which was destroyed decades ago, using this technique one can reconstruct a virtual 3 dimensional model of that place and can learn more about it. So, this way it can be used for teaching purposes in architectural history, while gaining more knowledge about the object.

This technique can be used in the fast-growing field of virtual reality also, for example till now 3-d models of an object have been created manually which can have manual errors. Using 3D reconstruction one can just take few images of an object and create a 3d model of the same which will be more accurate.

Another Application can be for medical purposes, like till now doctors use X-ray images of a person to detect any kind of disease or problem. Using X-rays taken from different directions it might be possible to create a real 3-D model which can help in locating a problem easily and more precisely. So, just by using few images one can land up with a 3-d model which can tell about the real structure of the entity under consideration.

II. RELATED WORK

A. Feature Matching

The earlier work for feature detection and matching was based on Moravec [8] work of utilizing corners as features points. This was improved by Forst[3] , Harris and Stephen[4] by incorporating gradient images to create cornerness function. Though these feature detection algorithms are translation and rotation invariant, they are susceptible to change of scale and application of affine transforms. This scale invariance was attained using SIFT proposed by lowe [7]. The limitation of SIFT is the time taken to compute features across multiple images. To address this, SURF was introduced by Bay [2] which incorporates the determinant of Hessian matrix to generate keypoints.

The feature matching techniques associated with these algorithm relies on matching patches around detected keypoints.

B. Structure from Motion

Structure from Motion is a collection of techniques with an objective to construct 3D scene, camera pose from a set of correspondence points. The earliest work used two-frame relative orientation presented by Longuet-Higgins [6]. This was extended to multi-frame technique using factorization methods by Tomasi-Kanade [20] and global optimization by Spetsakis-Aloimonos [16], Szeliski-Kang [18], Oliensis[10]

Bundle Adjustment introduced by Triggs [21] as an optimization technique that minimizes the re-projection error by refining 3D point and intrinsic and extrinsic paramters of the camera, can be noted as the next advancement that has been applied extensively in the reconstruction pipeline.

In circumstances, where the reconstruction is being performed in uncalibrated setting, self calibration techniques described by Pollefeys[11], [12] have provided successful results. An alternative would be the use of metadata obtained from the EXIF tags of the image.

Amongst the different applications of structure from motion, some of the most popular includes Microsoft Photosynth, Photo-tourism[15], MIT City Scanning Project[19], 4D Cities Project[14], Stanford City Block Project [13] and UrbanScape Project [1]

Fig. 1: Input Image

III. METHODOLOGY

A. Dataset

The project uses fountain-P11 dataset[17] which comprises of 11 high resolution sequence images of a fountain that are taken at 15deg of each other. The benefit of the dataset is the availability of camera and projection matrices.

B. Feature Extraction and Matching

Since the number of images used for reconstruction maybe in thousands, matching all the images to each other is infeasible. Thus significant keypoints are identified in all the images using SURF. To find matches in the images, nearest neighbor algorithm is used. However, since the feature descriptor have high dimensionality, using nearest neighbor for every match is extremely expensive computationally. To address the problem, an approximate version of nearest neighbor as described in [9] is used.

Fig. 2: Keypoints Image

For smaller dataset, opencv's implementation of SURF is used with a brute force matcher rather than nearest neighbor matcher.

C. Estimating Fundamental Matrix

Though keypoint matching using feature descriptors is extremely handy, the outcome still have significant mismatches that need to be pruned before moving forward with the reconstruction. An additional constraint is applied to the matched keypoints by computing the fundamental matrix . Fundamental matrix relates corresponding points in a stereo image pair based on the equation

$$x^T * F * x' = 0 \tag{1}$$

where, x, x' are stereo image and F is the fundamental matrix.

Since degree of freedom of the fundamental matrix is 9, minimum 8 points correspondence are required to calculate it. 8 point algorithm [6] along with RANSAC is used to determine fundamental matrix. An auxiliary benefit of applying RANSAC is the computation of outliers which can be removed to reduce the number of mismatches.

D. Camera Matrix

The matrix represents intrinsic camera parameters. It is used to transform 3d camera coordinates to image coordinates and vice versa. The coefficients of the matrix are where,

$$\begin{vmatrix} f_x & \text{s} & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{vmatrix}$$

TABLE I: Camera matrix

f_x and f_y are focal length in x and y coordinates. x_0 and y_0 is the image center.

Focal length of the camera can be obtained either from 5 point algorithm [6] or using XIF tags in jpeg images. The focal length obtained is in mm and must be converted to pixels using the following formula
focal length in pixels = (image width in pixels) * (focal length in mm) / (CCD width in mm)
where, CCD width is part of metadata associated with an image. If CCD width is missing from the XIF tags, it can be computed from a camera database obtained from openMVG.

E. Essential Matrix

Essential matrix is similar to fundamental matrix that relates point correspondences in two stereo images. Essential matrix can be computed using the formula

$$E = K^T * F * K[5] \tag{2}$$

where K is the camera matrix and F is the fundamental matrix.

F. Decomposition in Rotation and Translation matrix

Essential matrix can be decomposed into rotation and translation matrix using Singular Vector Decomposition using homogenized keypoint.

$$E = U * \Sigma * V^* \tag{3}$$

where U, V^* are unitary matrix, Σ is diagonal matrix

$$Rotation - Matrix = U * W * V^* \tag{4}$$

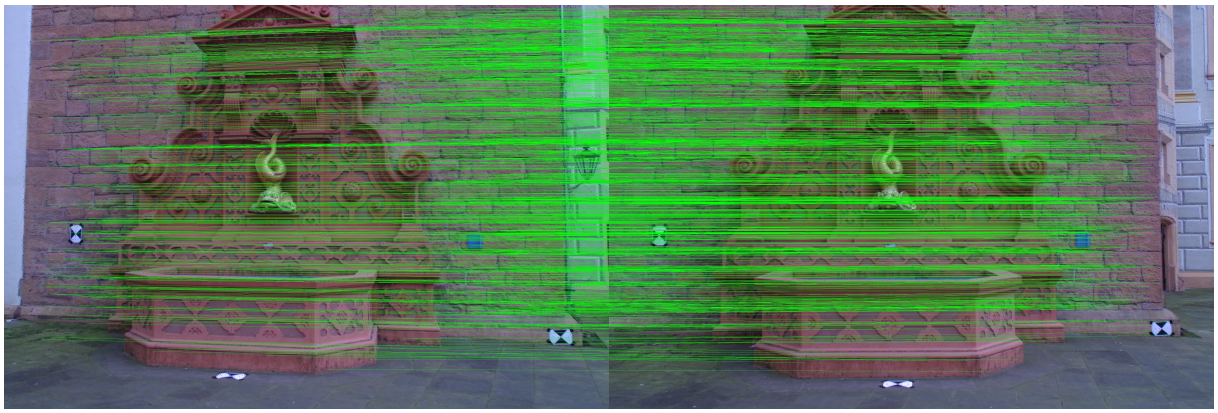

Fig. 3: Image Matches after Estimating fundamental Matrix

where W = $\begin{vmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{vmatrix}$. This results in four possibilities of rotation and translation matrix of which one projects the keypoints in front of the camera.

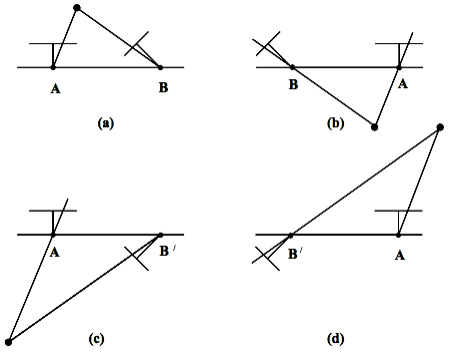

Fig. 4: Possibilities of the Rotation and Translation [5]

G. Triangulation

For every image pair, the first image lies at origin. Thus the rotation matrix is $\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{vmatrix}$ and translation vector is $\begin{vmatrix} 0 & 0 & 0 \end{vmatrix}$. The rotation and translation matrix of the second image was determined in the previous section.

The coordinates of a point in world coordinate space is determined based on the equation.

$$\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix}^{-1} s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{5}$$

where, (X,Y,Z) are coordinates of a 3D point in world coordinate space, (u,v) are coordinate of pixel in image, s is the scale coefficient of an image.

H. Triangulation Across Multiple Image

Triangulation using essential matrix suffers from scale ambiguity, thus reconstructed points computed across multiple images cannot be combined directly. This is formulated as Perspective-n-point problem where where camera pose must be estimated based on the 3D world coordinates and corresponding 2D image coordinates. The keypoints from the first image pair are used to start the triangulation. Every image is added using keypoints that have already been used in reconstruction to compute the new projection matrix. This projection matrix is used to triangulate new three dimensional points, thereby incrementally adding new 3D-2D point correspondences.

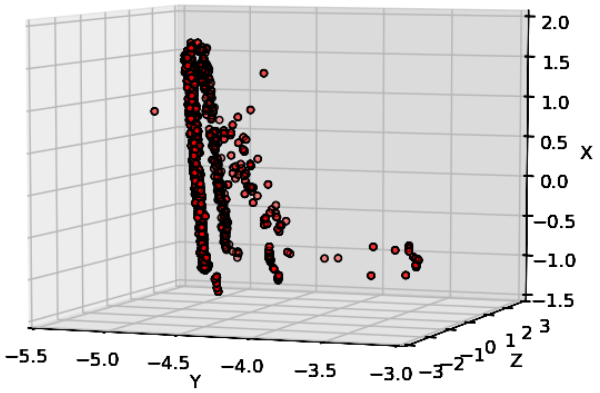

Fig. 5: Triangulation using 2 Images

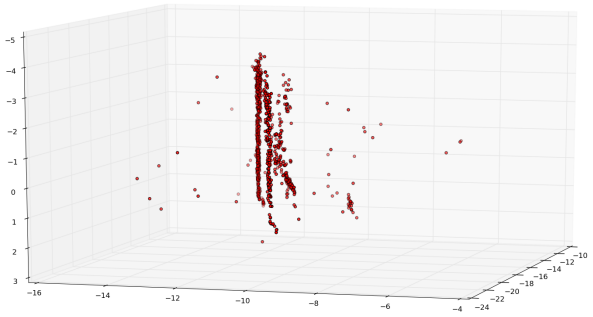

Fig. 6: Triangulation using all images

IV. RESULT

Fountain-P11 dataset was used to generate sparse model of the fountain using first the two images of the front of the fountain and then all images provided by the dataset. The second reconstruction adds more points to the first, though mismatches that might be present after pruning results in peculiar projection of a few points. These points could have been avoided by setting the value of determinant of Hessian matrix higher than current default or to make the criteria of RANSAC for more stringent but these results in a more sparser model.

V. DISCUSSION

In this project, we tried to implement and replicate the finding of the paper titled Photo-Tourism[15] at a smaller scale using Fountain-P11 dataset. Most of the implementation used the in-built function provided by opencv. Though we successfully finished reconstruction using 2 images, we were unable to extend it to multiple image. The most likely cause would be scale ambiguity when applying PnPRansac to determine the projection matrix for adding a new image.

VI. FUTURE WORK

Since this was a semester project, there were many aspects of reconstruction that we were unable to study or implement.

Some of the thing that we would like to add to the projects are:

- Extend the current SFM implementation to successfully work for multiple images
- Add Bundle adjustment to various stages of the algorithm to minimize re-projection error.
- Try and implement the current methods from scratch.
- Run the program for a very large dataset as described in the paper Photo tourism [15]

VII. ACKNOWLEDGEMENT

This work was submitted as a course project for CSCI-631 Introduction to Computer Vision. We are grateful to Prof. Srinivas Sridharan for his encouragement and helpful comments during different stages of implementation.

REFERENCES

[1] AKBARZADEH, A., FRAHM, J.-M., MORDOHAI, P., CLIPP, B., ENGELS, C., GALLUP, D., MERRELL, P., PHELPS, M., SINHA, S., TALTON, B., ET AL. Towards urban 3d reconstruction from video. In *3D Data Processing, Visualization, and Transmission, Third International Symposium on* (2006), IEEE, pp. 1–8.

[2] BAY, H., TUYTELAARS, T., AND VAN GOOL, L. Surf: Speeded up robust features. In *Computer vision–ECCV 2006*. Springer, 2006, pp. 404–417.

[3] FÖRSTNER, W. A feature based correspondence algorithm for image matching. *International Archives of Photogrammetry and Remote Sensing 26*, 3 (1986), 150–166.

[4] HARRIS, C., AND STEPHENS, M. A combined corner and edge detector. In *Alvey vision conference* (1988), vol. 15, Citeseer, p. 50.

[5] HARTLEY, R., AND ZISSERMAN, A. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[6] LONGUET-HIGGINS, H. C. A computer algorithm for reconstructing a scene from two projections. *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms, MA Fischler and O. Firschein, eds* (1987), 61–62.

[7] LOWE, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision 60*, 2 (2004), 91–110.

[8] MORAVEC, H. P. *The Stanford cart and the CMU rover*. Springer, 1990.

[9] MUJA, M., AND LOWE, D. G. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1) 2* (2009), 331–340.

[10] OLIENSIS, J. A multi-frame structure-from-motion algorithm under perspective projection. *International Journal of Computer Vision 34*, 2-3 (1999), 163–192.

[11] POLLEFEYS, M., AND GOOL, L. V. From images to 3d models. *Communications of the ACM 45*, 7 (2002), 50–55.

[12] POLLEFEYS, M., VAN GOOL, L., VERGAUWEN, M., VERBIEST, F., CORNELIS, K., TOPS, J., AND KOCH, R. Visual modeling with a hand-held camera. *International Journal of Computer Vision 59*, 3 (2004), 207–232.

[13] ROMAN, A., GARG, G., AND LEVOY, M. Interactive design of multi-perspective images for visualizing urban landscapes. In *Proceedings of the conference on Visualization'04* (2004), IEEE Computer Society, pp. 537–544.

[14] SCHINDLER, G., DELLAERT, F., AND KANG, S. B. Inferring temporal order of images from 3d structure. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (2007), IEEE, pp. 1–7.

[15] SNAVELY, N., SEITZ, S. M., AND SZELISKI, R. Photo tourism: exploring photo collections in 3d. In *ACM transactions on graphics (TOG)* (2006), vol. 25, ACM, pp. 835–846.

[16] SPETSAKIS, M., AND ALOIMONOS, J. Y. A multi-frame approach to visual motion perception. *International Journal of Computer Vision 6*, 3 (1991), 245–255.

[17] STRECHA, C., VON HANSEN, W., GOOL, L. V., FUA, P., AND THOENNESSEN, U. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (2008), Ieee, pp. 1–8.

[18] SZELISKI, R., AND KANG, S. B. Recovering 3d shape and motion from image streams using nonlinear least squares. *Journal of Visual Communication and Image Representation 5*, 1 (1994), 10–28.

[19] TELLER, S., ANTONE, M., BODNAR, Z., BOSSE, M., COORG, S., JETHWA, M., AND MASTER, N. Calibrated, registered images of an extended urban area. *International journal of computer vision 53*, 1 (2003), 93–107.

[20] TOMASI, C., AND KANADE, T. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision 9*, 2 (1992), 137–154.

[21] TRIGGS, B., MCLAUCHLAN, P. F., HARTLEY, R. I., AND FITZGIBBON, A. W. Bundle adjustmenta modern synthesis. In *Vision algorithms: theory and practice*. Springer, 1999, pp. 298–372.