ABSTRACT

In this project report, an approach to building visual models from camera images is described. These models can be used for graphics and virtual reality applications. The basic aim is reconstructing a structure that has been captured by any video camera—hand held or controlled motion, without having any apriori information about its intrinsic parameters like its focal length. The process begins by obtaining interest points from the various selected image frames, extracted from the video sequence taken, followed by matching these features to obtain relations between multiple views. From this, the structure and motion parameters can be retrieved. Self-calibration techniques can be employed to restrict the reconstruction ambiguity from a projective to a metric reconstruction. The analysis was done on not more than 3 images at a time (no batch update). Thus, the topic should be more appropriately redefined as '3D reconstruction from two images / a triad of images'.

Keywords: 3D reconstruction, hand-held camera, Structure-from-Motion, Projective reconstruction, Self-calibration, stereo matching.
1. INTRODUCTION & PAST WORK

The aim of this project was to put to use a subset of the various methods that have been proposed in the last decade and a half and see how one particular combination of these methods fair when used on personal data set. Substantial amount of work has been put into this field of computer vision covering various aspects of:

- intrinsic parameters of camera such as focal length: camera calibration-not a necessity now
- reconstruction from multiple images[1]
- scene reconstruction using Voxel colouring[2]
- novel view generation without going through reconstruction[3]
- reconstruction of non rigid objects like the human face[4]
- Modelling a miniature object on a turn table when the rotation angles & distance to the rotation axis are not accurately known[5]
- Affine factorization method to extract 3D from image sequences assuming orthographic projection.[6]
- and much more...

However, this project has been greatly motivated by the works of Fitzgibbon[7], Zisserman[8], Pollefeys[9] and a comprehensive report on 3D reconstruction prepared by Dang Trung Kien[10].

The basic outline of the 3D reconstruction procedure can be graphically shown as:

The building of interest was the HR Kadim Diwan Building- Department of Computer Science and Engineering at the Indian Institute of Technology, Kanpur. Of the various steps listed above, the project was successful up to the step of upgrading the projective reconstruction to a quasi-affine reconstruction. However, plane fitting, texturing and VRML model construction could not be applied to the dense point cloud obtained.

2. METHODOLOGY EMPLOYED

2.1 CAMERA CAPTURE:
Circumnavigation and capture of the building decided upon in a video using a normal camera (a camcorder with no aprori knowledge of its intrinsic parameters) courtesy SFS: Student’s Film Society at IIT Kanpur.

2.2 PRE-PROCESSING:
The video sequence is then preprocessed by selecting relevant noise-free frames, normalizing the illumination, etc.

2.3 FEATURE DETECTION:
To extract the interest points from the selected frames, Harris Corner Detection algorithm[11] was employed which essentially uses weighted Sum of Square Difference-SSD between any two image patches. Consider an image patch over the area (u,v) and shifting it by (x,y) gives SSD- S(x,y) as:

\[ S(x, y) = \sum_{u} \sum_{v} w(u,v) (I(u,v) - I(u-x, v-y))^2 \]


2.4 FEATURE MATCHING:
This is done to obtain correspondences between the features (Harris corners) extracted in the two selected frames. Matching is done by correlation- detect those points that are maximally correlated with each other within a window surrounding each point. Let the search window be (Dx,Dy). And (x,y) ranges over pixels in the block centered at (x0,y0). Then cross-correlation is given by:

\[ \text{Correlation}(D_x, D_y) = \sum_{(x,y)} I(x, y, t)I(x + D_x, y + D_y, t + D_t) . \]

Image Source: Book Pg#875 Russel, Norvig- Artificial Intelligence- A Modern Approach
2.5 EXTRACTING THE GEOMETRIC CONSTRAINTS:

2.5.1 FUNDAMENTAL MATRIX:

So as to determine the geometric constraints between the different views, Fundamental Matrix is derived using RANSAC (Random Sample Consensus).

Let \( x \) and \( y \) be the corresponding images of the 3D point \( X \) in the two views. Let \( z = y^\top \) transpose. Then \( F \) is the 3X3 fundamental matrix such that the following constraint is met:

\[ zFx = 0 \]

The algorithm used to automatically estimate the fundamental matrix between two images using RANSAC was obtained from Page# 275 Algorithm 10.4 of the book- 'Multiple View Geometry in Computer vision' by Richard Hartley and Andrew Zisserman, CAMBRIDGE publications.

2.5.2 EPIPOLAR GEOMETRY:

The epipolar lines and the epipoles determine the position of a 3D point from its projections in two images.

The epipolar lines \((L_1, L_2)\), the epipoles \((e_1, e_2)\), the fundamental matrix \(F\), its transpose: \(G\) and the image points \((x_1, x_2)\) are related as follows:

\[
\begin{align*}
L_2 &= F^*x_1 \quad ; \\
L_1 &= G^*x_2 \quad ; \\
F^*e_1 &= 0 \quad ; \\
G^*e_2 &= 0
\end{align*}
\]

THE EPIPOLAR CONSTRAINT  A point in one image generates a line in the other on which its corresponding point must lie. The search for correspondences is thus reduced from a region to a line. This constraint arises because, for image points corresponding to the same 3D point, the image points, 3D point and optical centres are coplanar.

[Quoted Text and Image 1 Source: CVonline<http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/EPSRC_SSAZ/node18.html>]

[Epipolar geometry (lines and epipoles) in 2 CSE frames]
2.6 OBTAINING THE CAMERA MATRICES:

The 3x4 projection / camera matrices project a 4x1 homogeneous 3D point (X) onto different 2D images as 3x1 homogeneous 2D poins (x) as

\[ x = PX \]

A pair of camera matrices \( <P_1,P_2> \) uniquely determine the fundamental matrix (F) but the reverse is not true. That is to say, the fundamental matrix determines the pair of camera matrices at best upto a right multiplication by a projective transformation. Thus a notion of canonical cameras comes into picture. A canonical form for pair of camera matrices \( <P_1,P_2> \) can thus be obtained such that \( <P_1*H , P_2*H> \) also yields the same F, where H is a 3D projective transformation (4X4).

The canonical cameras can be very easily determined once the fundamental matrix constraint and epipolar geometry has been derived, as follows:

\[ P_1 = [I | 0] = [ 1 0 0 0 \\
0 1 0 0 \\
0 0 1 0 ] \]

\[ P_2 = [M | m] = [-e2T*F | e2] \]

where: \( e2T = e2 \) transpose

\[ e2 = m \]

\[ e1 = inverse(M)*m \]

2.7 OBTAINING THE 3D POINTS:

Using singular value decomposition (SVD), the 3D points are extracted from the correspondences in the two images as follows:

\[ x_1 = P_1*X \quad ; \quad x_2 = P_2*X \]

\[ x_1 ** (P_1*X)=0 \quad ; \quad x_2 ** (P_2*X)=0 \]

where ** refers to cross product.

This yields a set of equations of the form AX=0

Using SVD on A gives X.

The following steps could not be incorporated in this project but play an essential part in the proper visualization of the final 3D reconstruction.
2.8 AUTO OR SELF CALIBRATION:

Converting the the projective structure (the canonical camera matrices obtained earlier) to a metric one for a proper visualization of the 3D structure. The crux here is that the projective structure obtained above has the following constraints many of which are quite obviously are not desired:

For a Projective Transformation:  
- Parallelism not maintained.
- Ratio of lengths or areas not preserved.
- A square may get converted to any quadrilateral.
- Although, concurrency and collinearity are preserved.

The following figure explains the point. What we want is a metric (similarity) reconstruction.

![Diagram of transformations from projective to metric]


2.9 VRML MODEL RECONSTRUCTION (Virtual Reality Modeling Language):

To the crude points structure, plane fitting and texturing is required to be able to visualize a 3D structure for the building of interest.
3. EXPERIMENTAL RESULTS

3.1 TWO IMAGES

3.1.1 FRAME SELECTION (CSE HR KADIM DIWAN BUILDING) ANGLE 1

3.1.2 HARRIS CORNER DETECTION

3.1.3 CORRELATION MATCHING VS RANSAC FILTERING
3.2 THREE IMAGES

3.2.1 FRAME SELECTION (CSE HR KADIM DIWAN BUILDING) ANGLE 2

3.2.2 HARRIS CORNER DETECTION

3.2.3 RANSAC FILTERING ON CORRELATION MATCHING
3.3 3D POINT CLOUD FOR THE CSE 3 IMAGE SET

4. CONCLUSION

As far as this project goes, upgrading the 3D point cloud obtained as above to a metric reconstruction, followed by plane fitting and texturing- VRML Model Construction should yield the expected result- the 3D reconstruction of the CSE building (angle 1,2). However, this could not be tested in this project.

In this project, the reconstruction was from a triad of images. This shall lead to an incomplete model since the views do not fully cover the object of interest. However, using the same techniques combined with batch update to incorporate plenty more frames would lead to a full 3D reconstruction and the elimination of the requirement of apriori knowledge of the camera's intrinsic parameters by self(auto) calibration techniques make this and related procedures quite useful for 3D modelling for graphics and virtual reality applications- 3D games, Virtual Worlds etc.
5. ACKNOWLEDGEMENT

I thank Prof. Amitabha Mukerjee for nurturing the basic concepts of Artificial Intelligence, particularly those relating to Computer Vision, in me and for his guidance which helped me a long way during the course of this project. A special thanks again to Sir for providing me with the book by Zisserman and Hartley - 'Multiple View Geometry in Computer Vision' which too guided me at every step.

6. REFERENCES

1. Multilevel modelling and rendering of architectural scenes Subhashis Banerjee et. al. IIT Delhi EUROGRAPHICS 2003, Volume 22 (2003), Number 3
3. Novel View Synthesis by Cascading Trilinear Tensors Shai Avidan and Amnon Shashua-Hebrew University, IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 4, NO. 4, 1998
4. Recovering Non-Rigid 3D Shape from Image Streams Christoph Bregler, Aaron Hertzmann, Henning Biermann New York University, Computer Vision and Pattern Recognition, 2000
5. Automatic 3D Model Construction for Turn-Table Sequences Andrew W. Fitzgibbon, Geo Cross and Andrew Zisserman, University of Oxford, Proceedings of the European Workshop on 3D Structure from Multiple Images of Large-Scale Environments, 1998
10. "A REVIEW OF 3D RECONSTRUCTION FROM VIDEO SEQUENCES" by Dang Trung Kien,Computer Science Deptt, University of Amsterdam. ISIS technical report series-2005