International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.102.29

Camera Position Estimation using 2D Image Dataset

Dr. Surita Maini[1], Dr.. Ashwani Kumar Aggarwal[2]
[1,2] *Department of Electrical and Instrumentation Engineering, Sant Longowal Institute of Engineering and Technology, Longowal, Punjab, India*

Abstract- In this paper, a method for estimating the position and orientation of a camera is proposed. The method works by use of interest points in frames of a digital video. In the first step, digital video is split into a number of keyframes. The keyframes are extracted based on calculation of motion vectors in consecutive frames. As the number of interest points is large in each frame, their pruning is carried out by use of local and global histograms. Feature vectors are calculated for each interest point which are based on several features of image. Among many features, we used shape features, texture features and variability in intensity levels. The distance between the feature vectors is calculated based on Euclidean distance. Due to camera noise and environment noise, outliers might be present which are removed using RANSAC. Camera position and orientation is estimated using epipolar geometry. As the camera potion is found with interest points in a sparse fashion, interpolation is used to find smooth camera path. To evaluate the performance of the method, several types of video including broadcast video and sports video is used. It is observed that proposed method performs better than existing methods used for camera position estimation.
Keywords –Interpolation, Indices, Epipolar , Camera

I. INTRODUCTION

Camera position and orientation need to be known for many computer vision applications such as 3D reconstruction, video mosaicking, image stitching etc. Structure from motion methods give camera position and orientation along with 3D reconstruction of the scene [1]. Such methods work when changes in camera pose in consecutive frames is small. Simultaneous Location And Mapping (SLAM) methods give camera pose while simultaneously constructing 3D scene [2]. A digital video is a sequence of frames each of which is a two dimensional array representing the spatial distribution of intensity values at particular time. Motion in 3D is casted as motion vector in two consecutive frames of a video sequence.

Among many methods, camera pose estimation is achieved using Perspective-n-Point (PnP) method for aerial video in [3]. The method uses 2D-3D matching for camera pose estimation. Point cloud constructed from omnidirectional images is used for camera pose estimation [4]. Camera pose is estimated in [5] using CAD model of any object present in a scene. The entire 3D map of the scene need not be known for the pose estimation. Single ground level omnidirectional image along with 2D. The map of the buildings is used to estimate camera pose in [6]. All these methods either use 3D map of the scene or rely on CAD methods for estimating camera pose in a video. A simple yet efficient method based on interest point detection is presented below.

II. KEY FRAME EXTRACTION

2.1 Digital Video Sequence

Digital video is sequence of frames and each of which is a RGB image with bit depth of 8-bits. A 256×256 digital frame is $256\times256\times8\times3$ bit image. A digital frame is obtained from an analog picture with the process of 2D sampling.

$$I(m,n)=A(x,y)\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\delta(x-mX_s,y-nY_s) \tag{1}$$

In case of two-dimensional image, after a DWT transform, the image is divided into four corners, upper left corner of the original image, lower left corner of the vertical details, upper right corner of the horizontal details, lower right corner of the component of the original image detail (high frequency). You can then continue to the low frequency components of the same upper left corner of the 2nd, 3rd inferior wavelet transform.

Let frame f_i be i^{th} frame of a video consisting of N frames. The complete video sequence is represented as below.

$$F=f_i \quad \forall i=1,2,...,N \tag{2}$$

Two consecutive frames of a video sequence are shown in Figure 1.

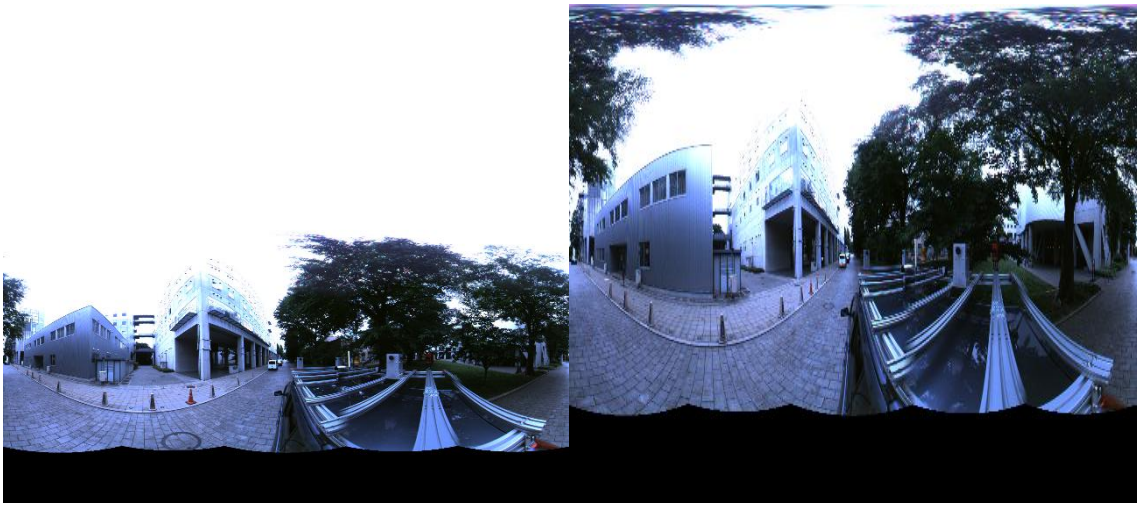

Figure 1. Video frame sequence

2.2 Extraction of Key Frame –

Extraction of key frame involves process of extracting representative frames in a digital video where sudden change in content takes place. Key frame extraction is important in many computer vision applications such as video indexing and image retrieval. DCT coefficients are used in feature extraction step in [7]. Following methods are used for key frame extraction.

2.3 Average Pixel Method

In this step, shot boundary is detected. Consider there are N_k^S frames in shot k. Then average pixel $A(i, j)$ for shot k is given as below.

$$A(i, j) = \frac{1}{N_k^S} \sum_{1}^{S_k} p(i, j) \tag{3}$$

Average fame of shot k is computed for each pixel in the frame. The distance of each frame in shot k from the average frame is computed. A frame with minimum distance is selected as key frame for shot k.

2.4 Average Histogram Method

In this step also shot boundary is detected. For each frame in a shot, normalized histogram of grey level intensity is calculated.

$$p_r(r_k) = \frac{n_k}{N} \quad 0 \le r_k \le 1, \quad k = 0,1,2,\ldots,L-1 \tag{4}$$

where r_k is normalized intensity level value.

L is number of gray scale levels in the image

n_k is no. of pixels with gray level r_k

and N is total number of pixels.

Average frequency of pixels in each bin of histogram is calculated. Average histogram of all shots is obtained and is normalized. A key frame is selected where normalized histogram and average histogram are similar.

For this purpose the following formulae is use-

$W(i) = (y_w(i) + y_o(i)) / \alpha$-2

After this Execution the Inverse 2-level discrete wavelet transform is applied on the watermark data to generate three watermark images extracted.

III. INTEREST POINT DETECTION

Template based image matching suffers from drawbacks of image size difference, rotation and other transformations in the image. To match two images, interest points are detected in the two images. The interest points are chosen in such a way that these points are robust and repeatable. The interest points are chosen to be scale invariant and robust to illumination changes.

A number of interest points detected on a spherical image are shown in Figure 2. If the image patch contains large texture, denser interest points are obtained.

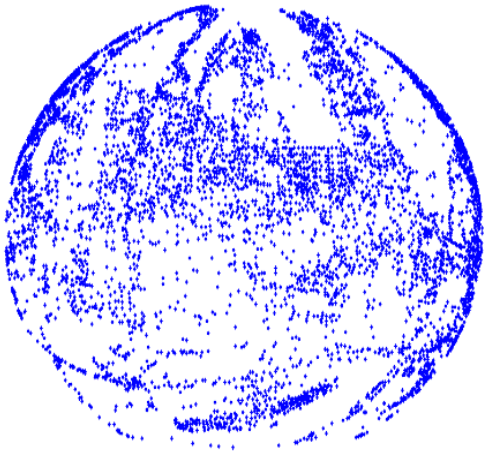

Figure 2. Watermark embedding algorithm Block Diagram

IV. DESCRIPTOR MATCHING

Around each interest point extracted from the image, a descriptor representing the environment around interest point is calculated. Among many, SIFT descriptors are used widely which is of size 128×1[8]. Descriptors in two images are compared using distance metric. Any of the following distance metrics are used.

Euclidean distance

This distance is most widely used distance metric. Euclidean distance finds the distance between vectors of feature vectors of two images [9].

$$d = \sum_{i=0}^{n} \left| I_i^1 - I_i^2 \right|^2 \tag{5}$$

Canberra distance

The feature vector distance is normalized by dividing the distance with sum of feature vectors magnitudes.

$$d = \sum_{i=0}^{n} \frac{\left| I_i^1 - I_i^2 \right|}{\left| I_i^1 \right| + \left| I_i^2 \right|} \tag{6}$$

Sum of Squared absolute distance (SSAD)

This distance is sum of squares of difference between magnitudes of feature vectors of two images.

$$d = \sum_{i=0}^{n} \left(\left| I_i^1 \right| - \left| I_i^2 \right| \right)^2 \tag{7}$$

Sum of absolute distance(SAD)

This distance calculates sum of difference of absolute value of feature vectors of two images.

$$d = \sum_{i=0}^{n} \left| I_i^1 \right| - \left| I_i^2 \right| \tag{8}$$

Maximum value distance

This distance is used to calculate the largest value of distance between feature vectors of two images.

$$d = \max\left(\left| I_1^1 - I_1^2 \right|, \left| I_2^1 - I_2^2 \right|, ..., \left| I_n^1 - I_n^2 \right| \right) \tag{9}$$

In descriptor matching, first match and second match is calculated. A match is discarded with the distance ratio of first match to second match is more than 0.8.

V. CAMERA POSE CALCULATION AND INTERPOLATION

A feature in an image is projection of 3D point in space as shown in Figure 3.

http://dx.doi.org/10.21172/ijiet.102.29

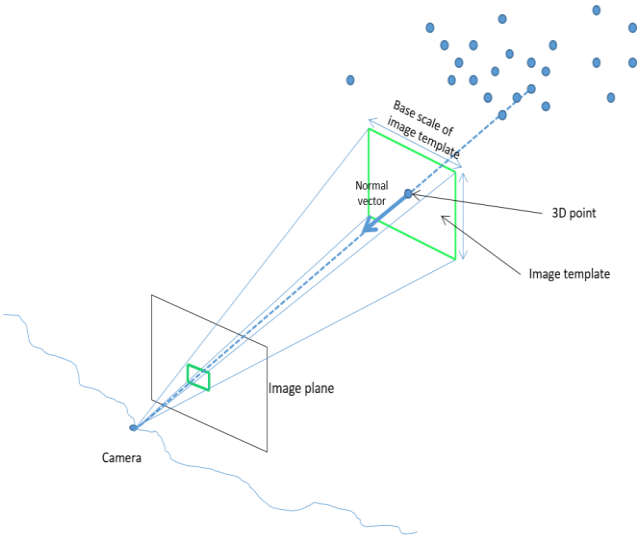

Fig.3. Projection of 3D point on image plane

The relationship of pixel in an image and its corresponding 3D point is given by Epipolar geometry. A 3D point after 3D rotation and translation is mapped to another 3D point as given below in equation (10).

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = R \begin{bmatrix} x \\ y \\ z \end{bmatrix} + T \tag{10}$$

Where $R_{3\times3}$ is rotation matrix and $T_{3\times1}$ is translation matrix.

From the pixels in the image and the 3D point, camera pose is estimated in the key frames. Camera pose in the whole video is obtained by interpolation of camera pose obtained in key frames of the complete video sequence.

VI. DESCRIPTOR MATCHING

This paper presents an algorithm for camera pose estimation in digital video using points of interest. This method works by matching descriptors in key frames and obtaining camera pose in these key frames.

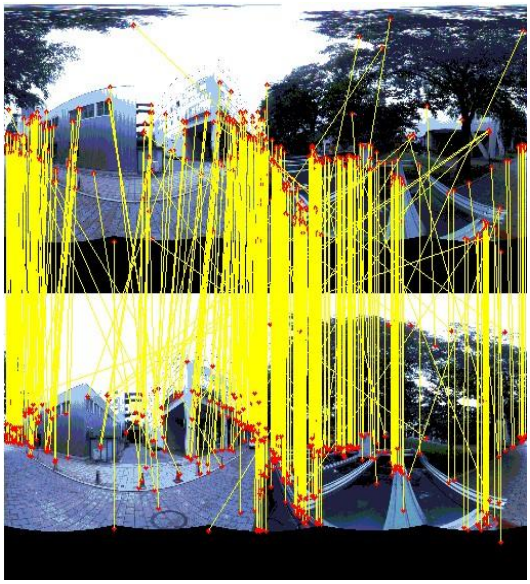

Fig.4. Descriptor matching

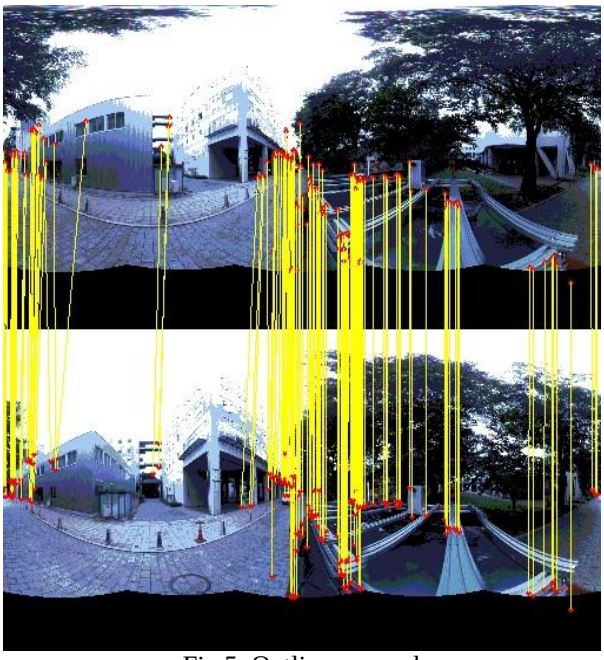

Fig.5. Outlier removal

The complete camera pose estimation is achieved by interpolating the camera pose in consecutive frames. Figure 4. shows interest point matching and Figure 5. shows matchings after removing outliers using RANSAC.

VII. REFERENCE

[1] M.J. Westoby, J. Brasington, N.F. Glasser, M.J. Hambrey, J.M. Reynolds, 'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications", Geomorphology, Volume 179, 15 December 2012, pp 300-314.

[2] L. Zhao, S. Huang and G. Dissanayake, "Linear SLAM: A linear solution to the feature-based and pose graph SLAM based on submap joining, Proc. of *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Tokyo, 2013, pp. 24-30.

[3] Z. Kang and G. Medioni, "3D Urban Reconstruction from Wide Area Aerial Surveillance Video", *Proc. of IEEE Winter Workshops on Applications and Computer Vision (WACVW)*, Waikoloa, HI, 2015, pp. 28-35.

[4] J. Ventura and T. Höllerer, "Wide-area scene mapping for mobile visual tracking," *Proc. of IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Atlanta, GA, 2012, pp. 3-12.

[5] G. Bleser, H. Wuest and D. Stricker, "Online camera pose estimation in partially known and dynamic scenes", Proc. of *IEEE/ACM International Symposium on Mixed and Augmented Reality, ISMAR 2006*, Santa Barbard, CA, pp. 56-65.

[6] T. J. Cham, A. Ciptadi, W. C. Tan, M. T. Pham and L. T. Chia, "Estimating camera pose from a single urban ground-view omnidirectional image and a 2D building outline map", Proc. of *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, 2010, pp. 366-373.

[7] M. Chatzigiorgaki and A. N. Skodras, "Real-time keyframe extraction towards video content identification", *Proc. of 16th International Conference on Digital Signal Processing*, Santorini-Hellas, 2009, pp. 1-6.

[8] Baoming Shan, Fengying Cui, "Image Matching Based on Local Invariant Feature and Histogram-Based Similar Distance", Proc. of First International Workshop on Education Technology and Computer Science. ETCS '09., vol. 1, pp. 1030-1033

[9] Liwei Wang, Yan Zhang and Jufu Feng, "On the Euclidean distance of images," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, Aug. 2005, pp. 1334-1339.