POSE ESTIMATION OF IMAGE SEQUENCE CAPTURED FROM URBAN ENVIRONMENT

M. Mazaheri, M. Saadatseresht

Dept. of Geomatic Engineering, The University of Tehran, Iran - (mazaherim, msaadat)@ut.ac.ir

Commission III, WG III/5

KEY WORDS: Pose Estimation, Urban Environment, Structure from Motion, Feature Extraction, Long Sequence

ABSTRACT:

This paper reviews whole process of long sequence pose estimation and its challenges for urban environment which is considered as complex scene due to moving objects, greens, repetitious features and scene incidental events. In this study, color images are used for point feature selection and green features are filtered using hue values of surrounding window. Pyramidal KLT color feature tracker is used in short and wide base line fashion to investigate their behaviour in long sequence. To detect outliers, RANSAC in conjunction with common relative orientation algorithms (eight point and Nister five point) and space resection is used to compare their performance. Sequential local bundle adjustment estimates approximate camera poses and global bundle adjustment is used to refine the result.

1. INTRODUCTION

Structure from Motion (SfM) is one of the classical problems in field of computer vision which is aimed to reconstruction by analyzing relative object-camera motion in image sequences. The most important and complicated part of SfM is to find movement of features in the image sequence to extract camera pose. Urban environment is a complicated scene due to wide variation of object shape, moving and repetitious objects, greens and scene incidental events which causes to noise and outliers in feature extraction and matching process. Hence, robust algorithms should be used in every stage of the process.
In this section we review whole process of image sequence pose estimation including feature selection, short and wide base line feature tracking, key frame selection and outlier detection. Detail of our implementation is described in section 2 and the result of study is shown in section 3.

1.1 Feature Selection

Points are one of the easiest features to detect in images by means of differentiation. The most necessary characteristic of a point feature is to be distinguishable from its neighborhoods and stable to be found accurately in other images (Pollefeys, 2002).
Common criterion for feature matching is Sum of the Squared Differences (SSD) which supposed to be zero between correspondent features windows. According to SSD criterion, best features to track can be selected by structure matrix G (Bigün *et al.*, 1991):

$$G = \begin{bmatrix} \sum_w \overline{\mathbf{f}_x . \mathbf{f}_x} & \sum_w \overline{\mathbf{f}_x . \mathbf{f}_y} \\ \sum_w \overline{\mathbf{f}_x . \mathbf{f}_y} & \sum_w \overline{\mathbf{f}_y . \mathbf{f}_y} \end{bmatrix} \quad (1)$$

In which $\mathbf{f}$ is a multi-channel image, $\mathbf{f}_x$ and $\mathbf{f}_y$ are spatial derivatives and bar ($\bar{}$) indicates convolution with Gaussian filter and w is an appropriate window which must be large enough to increase discriminative power of the feature and small enough to decrease effects of image projective deformations (Kanade and Okutomi, 1991). Quality of features are defined by minimum Eigen value of G (Shi and Tomasi, 1994) or Harris criterion $R=\det(G)-k\times\mathrm{trace}(G)$ (Harris and Stephens, 1988). So, high quality features are the ones with most gradient variations in two independent directions, hence corner features are desired. “Choosing local maximum quality, make possible to expect stability in the selection process itself, so that windows selected in the next image are usually placed near the right position” (Ferruz and Ollero, 2000).
Simple thresholding of the quality values may results to remain features only in strong textured areas of the image. To spatially distribute the features, tiling idea can be used. A straightforward method is to sort features from high to low quality and starting from first feature, close features are removed according to the tiling cell size. This process has to be repeated for other features to obtain a purged feature list which is well spatially distributed.

1.2 Feature Tracking

Feature tracking is done in short and wide base line forms. For short base line feature tracking, the relation between images can be modelled with simple translational model $I_1(x)=I_2(x+d)$ in which I is illumination and d is displacement which is calculated iteratively by equation (2) to achieve sub pixel accuracy (Tomasi and Kanade, 1991):

$$b = \begin{bmatrix} I_x I_t \\ I_y I_t \end{bmatrix} \quad d = -G^{-1} b \quad (2)$$

Temporal derivative I_t can be approximated with I_2-I_1 in case of small displacement, otherwise, tracking process should be done in pyramidal fashion in which displacement is calculated at upper part of pyramid and scaled to finest level (Bouguet, 2000). In case of color image $\mathbf{f}=f(R, G, B)$, equation (2) provides three constraints on d (Barron and Klette, 2002).

Short base line feature tracking suffers from cumulative error and features may be missed after tracking in several frames. One solution is to perform a consistency check between tracked features with original one. It means to monitor similarity of image window around the original feature and its correspondence during the sequence. Figure 1 shows an example of wrong feature tracking because of pole movement in front of the door.

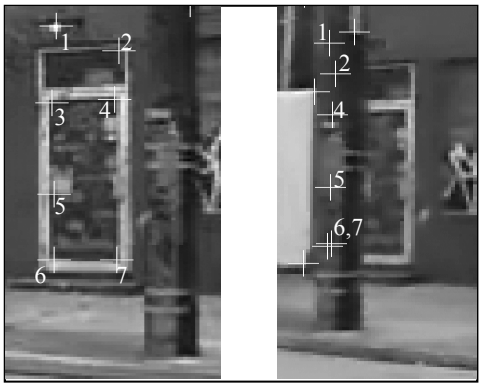

Figure 1: wrong correspondence due to short base line feature tracking (closer object passes farther one).

In an image sequence captured from urban environment, features are supposed to emerge and disappear in a several frame after, depending on platform speed and frame rate. Therefore, at each new image, feature selection is performed and those with specific distance to tracked ones are removed and the remaining is considered as new features.
Tracking of a feature may fail in immediately next image of the sequence because of scene incidental events. To enrich the matching process, Tracking of failed features can be done in neighbor images.
In wide base line tracking, displacement of features can no longer be modeled by translation model. A commonly adopted model is affine with illumination scale (λ) and shift (δ) parameters $I_1(x)=\lambda I_2(Ax+d)+\delta$. Affine model can be implemented in pyramidal fashion as well as translation model (Bouguet, 2001).
Another criterion is NCC score which is invariant to scale and shift in illumination and window size, therefore, it can be used in wide baseline feature matching. NCC score of all features in both of images is calculated and features are labeled as matched while they have maximum NCC score among other features in both sides.

1.3 Key Frame Selection

In case of high frame rate or slow platform speed, adjacent images are very close to each other which result in strong feature tracking but weak structure for reconstruction. The solution is to use key frames instead of whole sequence by analyzing the tracked features position and number. In general, key frames should be selected at suitable locations for reconstruction. The larger base line between key frames leads to stronger structure but number of matched features decreases instead. Three factors: (a) the ratio of the number of correspondent points (R_c) to total number of features (R_t), (b) the homography error and (c) spatial distribution of correspondent points; form a cost function to detect key frames (Seo *et al.*, 2003). (Pollefeys *et al.*, 2004) uses Geometric Robust Information Criterion (GRIC) (Torr *et al.*, 1998) which is related to goodness of fit; Key frame is selected once epipolar geometry explains the relationship between the pair of images better than homogeraphy model (GRIC(F) < GRIC(H)). (Ahmed *et al.*, 2010) propose an algorithm for robust key frame extraction using these criteria in addition to Point to Epipolar Line Criterion (PLEC).

1.4 Outlier Detection

As previously discussed, considerable amount of outliers happens in urban scenes and have to be detected with robust algorithms. RANSAC (Fischler and Bolles, 1981) is commonly used algorithm for outlier detection in field of computer vision. In two view geometry, RANSAC with relative orientation algorithms is used and outliers is defined by distance to epipolar line or Sampson error (Sampson, 1982) more than a given threshold.
In image sequence captured from urban environment (especially by car platform), epipolar lines are mostly in horizontal direction. Object repetition is more likely to happen in horizontal direction (e.g. window corners) and they are prone to be matched wrongly because of similar texture (Figure 2). This type of error cannot be detected by distance to epipolar error.

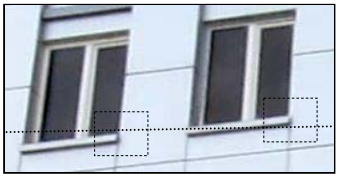

Figure 2: depicts two corner features with same texture and approximately equal distance to epipolar line.

Considering three views, it is possible to predict a position for a point in third view by intersecting two epipolar lines from first and second image (even in uncalibrated views). The drawback of this calculation is when epipolar lines are nearly parallel that cause to ambiguity in intersection point. This degenerate configuration happens frequently in urban scene image sequence. Fortunately, when center of projections are not coinciding (even they be collinear), a 3D point can be calculated using first and second views and project to third view to predict its image position. To detect outliers by means of 3D view geometry, pose of underlying image is estimated by resection upon reconstructed 3D points from previous images and outliers is defined by Euclidean distance of predicted point to observed point.
Previous studies used different algorithms with RANSAC such as Grunert (Mouragnon *et al.*, 2009), Bundle adjustment (Mayer, 2006) and five point (Nistér, 2004) to detect outliers via image or object space.

2. METHODOLOGY

To begin with, point features in color image is detected by Harris operator with direct channels gradients $\mathbf{f}_x=(R_x, G_x, B_x)$. Direct use of gradients extracts specular or shadow-shading points which may be undesired. A new class of derivative proposed by (van de Weijer *et al.*, 2005) can be applied to detect quasi invariant features which supposed to be insensitive to specular or shadow-shading.
To find a threshold for feature selection, we use Cumulative Distribution Function (CDF) of feature quality (minimum eigen

values) and select a value as threshold (t) where one percent of features can pass $P(X>t)$ =1%.
In urban scenes, trees produce many unstable features which cannot be tracked in other images and have to be filtered out from feature database. To filter such features, we use proportion of number of pixels between 60º - 120º hue value to total pixels inside the window more than a threshold (0.5 in our study). Since hue value is stable toward illumination changes, this filter works efficiently in different lighting situations (Figure 3).

Figure 3: detected features on a tree (left) after applying hue filter (right).

Since we work in color space, displacement (d) of features can be estimated by extending equation (1), (2):

$$\mathbf{G}=\left[G_R, G_G, G_B\right]^T, \mathbf{b}=\left[b_R, b_G, b_B\right]^T, d=\mathbf{G}^{-1}\mathbf{b} \qquad (3)$$

We select key frames according to value R_c/R_t and mean displacement of features to image width (d_f/d_w) once approaches to a given threshold (0.8 in our experiment).
Three matching strategies (a) short base line tracking (b) wide base line tracking and (c) NCC score matching is also investigated.
To detect outliers, we use common relative orientation algorithms for two view and resection for three view geometry to compare their performance. RANSAC with eight point algorithm is widely used because no camera calibration is needed. Nister five point algorithm (Nistér, 2004) linearly solves relative orientation problem and it is well suited for numeric implementation. RANSAC in conjunction with nonlinear resection is used here for outlier detection using three views.
To estimate pose of all views, two initial images are selected to define global coordinate system and must be matched strongly with possible larger baseline. Camera pose of initial images are estimated using coplanar relative orientation and initial 3D model is reconstructed by their correspondent points. A slipping window with three views is used for local bundle adjustment to optimize camera pose and 3D points. To begin with, two initial images with definite pose and new image points enter to bundle adjustment. Approximate camera pose of new image is estimated by resection upon initial 3D points or simply assumed equal to previous image. Bundle adjustment window slips on whole sequence and extracts camera poses and 3D coordinates of image points. A global bundle adjustment is done to refine whole parameters.

3. RESULTS

Before data capture, we calibrated the camera (Canon SX 200 IS) using several videos captured from a test field. Then, a HD video (1280×720 pixel resolution and 30 fps) was captured from urban environment on a car platform and the video (in MOV format) was spitted in full quality images which is easier to work with.
We applied three mentioned outlier detection methods (RANSAC with eight point, Nister five point and resection) and three strategies of feature matching (short and wide baseline tracking, NCC score matching) on this data set to compare their performance. Mean reprojection error of 3D points in to images (e) is considered as performance indicator according to below equation:

$$e=\frac{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}-\pi(X_i))^2}}{\sum_{i=1}^{n}v_i} \qquad (4)$$

In which x_{ij} is observed image coordinate of 3D point X_i in image j and v_i is number of images contain X_i. Function π projects 3D point into 2D image point.
Table 1 shows the mean reprojection errors (e) of different matching and outlier detection methods on the captured image sequence:

Matching method / Outlier detection method	Short base line	Wide base line	NCC
No outlier detection	0.0150	0.021	0.019
Eight point	0.0081	0.013	0.015
Nister five point	0.0073	0.011	0.011
Nonlinear resection	**0.0065**	0.010	0.009

Table 1: Mean reprojection error (e) by performing different outlier detection and feature matching strategies.

According to result, short base line feature tracking (with consistency check) and using RANSAC with resection as outlier detection method shows relatively better performance.
The values of e is not necessarily the best possible for each method because there exists several thresholds in all stage of the process which changes to each one, may cause to variation in final accuracy.

4. CONCLUSION

In this paper, we have reviewed process of pose estimation in long sequence and its key challenges. Feature selection and tracking in color images and filtering unstable green feature has been discussed.
Short and wide base line feature tracking has its benefits and disadvantages as previously mentioned. Using short base line feature tracking with consistency check shows better performance in long sequence pose estimation.
In two views, outliers are detected by distance to epipolar line criterion which is not enough constraint for errors that happen along epipolar line. RANSAC with resection upon 3D points can deal sufficiently with outliers.

REFERENCES

Ahmed, M., Dailey, M., Landabaso, J. & Herrero, N., 2010. Robust key frame extraction for 3d reconstruction from video streams. In: *Proceedings of VISAPP Conference*.

Barron, J. & Klette, R., 2002. Quantitative color optical flow. In: *Proceedings of 16th International Conference on Pattern Recognition*, pp. 251-255.

Bigün, J., Granlund, G. & Wiklund, J., 1991. Multidimensional orientation estimation with applications to texture analysis and optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 13 (8), pp. 775-790.

Bouguet ,J., 2000. Pyramidal implementation of the lucas kanade feature tracker description of the algorithm. *Microprocessor Research Labs, Intel Corporation.*

Bouguet, J., 2001. Pyramidal implementation of the affine lucas kanade feature tracker—description of the algorithm. Intel Corporation.

Ferruz, J. & Ollero, A., 2000. Real-time feature matching in image sequences for non-structured environments. Applications to vehicle guidance. *Journal of Intelligent and Robotic Systems,* 28 (1), pp. 85-123.

Fischler, M & .Bolles, R., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM,* 24 (6), pp. 381-395.

Harris, C. & Stephens, M., 1988. A combined corner and edge detector. In: *Proceedings of Alvey Conference*, pp. 189-192.

Kanade, T. & Okutomi, M., 1991. A stereo matching algorithm with an adaptive window: Theory andexperiment. In: *Proceedings of IEEE International Conference on Robotics and Automation*, pp. 1088-1095.

Mayer, H., 2006 of Conference. 3d reconstruction and visualization of urban scenes from uncalibrated wide-baseline image sequences. *In*: *IAPRS Volume XXXVI part 5*, pp. 207-212.

Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F. & Sayd, P., 2009. Generic and real-time structure from motion using local bundle adjustment. *Image and Vision Computing,* 27 (8), pp. 1178-1193.

Nistér, D., 2004. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 756-777.

Pollefeys, M., 2002. Visual 3d modeling from images-tutorial notes. University of North Carolina-Chapel Hill.

Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J. & Koch, R., 2004. Visual modeling with a hand-held camera. *International Journal of Computer Vision,*59 (3), pp. 207-232.

Sampson, P., 1982. Fitting conic sections to "very scattered" data: An iterative refinment of the bookstein algorithm. *Computer Graphics and Image Processing,* 18 (1), pp. 97-108.

Seo, J., Kim, S., Jho, C. & Hong, H., 2003. 3d estimation and keyframe selection for match move. In: *Proceedings of ITC-CSCC*.

Shi, J. & Tomasi, C., 1994 of Conference. Good features to track. *In*: ed. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593-600.

Tomasi, C. & Kanade, T., 1991. Detection and tracking of point features. Report CMU-CS-91-132, Carnegie Mellon University, Pittsburgh, P, USA.

Torr, P., Faugeras, O., Kanade, T., Hollinghurst, N., Lasenby, J., Sabin, M. & Fitzgibbon, A., 1998. Geometric motion segmentation and model selection. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences,* 356 (1740), pp. 1321-1340.

Van De Weijer, J., Gevers, T. & Geusebroek, J., 2005. Edge and corner detection by photometric quasi-invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 27 (4), pp. 625-630.

ACKNOWLEDGEMENT

Special thanks to National Elite Foundation of Iran for supporting M. Mazaheri to attend in PCV 2010.