COMBINING KALMAN FILTERING AND VISION-BASED TRAJECTORY ESTIMATION

Fuan TSAI[ab*], Huan CHANG[a], Yih-Shyih CHIOU[b], Yao-Tsung LIN[c] and Shin-Hui LI[c]
[a]Department of Civil Engineering, National Central University
[b]Center for Space and Remote Sensing Research, National Central University
[c]CECI Engineering Consultants, Inc., Taiwan

300 Jhongda Rd., Jhongli, Taoyuan County 32001, Taiwan
Tel: +886-3-4227151 ext. 57619; Fax: +886-3-4254908
Email: ftsai@csrsr.ncu.edu.tw; 1984chang@gmail.com; choice@csrsr.ncu.edu.tw;
tc506@ceci.com.tw; shl@ceci.com.tw

KEY WORDS: Homography, motion estimation, computer vision, visual navigation, normalized cross correlation

Abstract: The objective of this research is using video frames or highly overlapped images to obtain the camera orientation and translation parameters for trajectory estimation. One form of measurements comes from the computer vision community where successive frames from a camera approximately looking at the ground can be used to compute the translation between frames. In order to deal with the corner effect and registration problems, normalized cross correlation (NCC) is used to recognize the landmarks as the control points. The developed algorithms consist of four steps for camera trajectory estimation: (1) feature points detection and matching; (2) homography calculation; (3) control points detection and registration; (4) motion estimation. The first step is data decimation in order to reduce data amount and increase computation efficiency. Then, corner detector is employed to extract the feature points and match them between the frames using sum of absolute differences (SAD). The metric part of homography can provide the camera orientation and translation parameters according to the conjugate points between each frames. After that, this research uses RANSAC to remove the outlier of the previous step. NCC is then used to check if the camera pass thought the control points or not. This study compared the results with on-site measurement and with or without Kalman filter. Examples of applying the developed algorithm to tracking applications demonstrate the effectiveness of the methods. The example in outdoor environment indicates that the developed method for determining camera orientation and translation parameters can be used in providing initial conditions to real-time positioning and tracking in indoor or outdoor environments.

INTRODUCTION

Recently smartphone and pad have become more and more functional and popular. This type of equipments usually has a camera and can take video or photos with low cost micro electro mechanical systems (MEMS) sensor providing equipment's real-time attitudes. Some of them can provide acceleration information also. They can be used as references for outdoor GPS or indoor WiFi positioning to make real-time positioning more accurate. Using a single camera to obtain image features and to trace the feature by extend Kalman filter (EKF) can also track in unknown scenes to rebuild the path of camera movement. In order to increase the accuracy of the vision-based trajectory estimation using single camera, a vision-assisted scheme based on NCC approach is proposed to detect landmark locations as calibration technique. This technique is used to alleviate the corner-effect problem caused by filtering and tracking algorithms. Combining the inertial measurement unit (IMU) and WiFi signal to computer vision is an important issue. Therefore, this research uses computer vision to obtain the camera movement path for positioning and tracking applications.

1.1 Related works

Simultaneous localization and mapping (SLAM) systems have been presented which are capable of large-scale, accurate and real-time processing. Although most of SLAM system requires stereo vision, monocular SLAM is more challenging but conveniently to be installed and obtain data. Single camera has a limitation when measuring metric scale, so mapping monocular images must assume scale as an undetermined factor. Scale factor estimation can done by adding additional information source such as a calibration object (Davison, 2003) or by exploiting nonholonomic motion constraints (Scaramuzza et al., 2009). Strasdat et al. present an algorithm based on non-linear optimization which can dealing with drift effect in large scale scene (Strasdat,2010).

For this approach to work, feature points must be extracted effectively from overlapped images or video frames in order to match the images and subsequently to estimate the camera parameters. The feature point extraction must overcome various image defects, such as noise and blurs etc., thus requires sophisticated corner or feature detectors (Yalin and Yasemin, 2008). For the positioning and tracking, GPS is commonly used as a supplemental equipment to provide location information. However, if GPS is unavailable or the signal is low, other instrument such as IMU can also be helpful. An IMU usually has a shifting effect in time, so it must be corrected with feature points and camera parameters obtained from computer vision and photogrammetry. A possible solution is to use FAST corner detection to obtain features on every image or video frame. Then use sum of squared differences (SSD) to match detected features. After that uses BaySAC algorithm to remove the outliers to evaluate the camera's position (Hide et al., 2010). Then a single view movement evaluation algorithm similar to Bouguet and Perona (1995) can be used to estimate the camera's movement path and position immediately in unknown environment.

METHODOLOGY

This paper uses matching corresponding co-planar feature points of sequential video frames to obtain the translation and orientation parameters of the camera. The developed procedure for applying computer vision for motion estimation is displayed in Figure 1. First step is to decimate the video frame data and using corner detection to detect features in every image. Secondly, extracted feature points are matched using SAD algorithm. After preliminary image matching, RANSAC algorithm is applied to remove the outliers and calculate the shift parameters for a KF algorithm to estimate the camera's position. In order to recognize the reference nodes as control points along the test path, this paper combines NCC method with identifications of reference nodes for calibrating the location estimation.

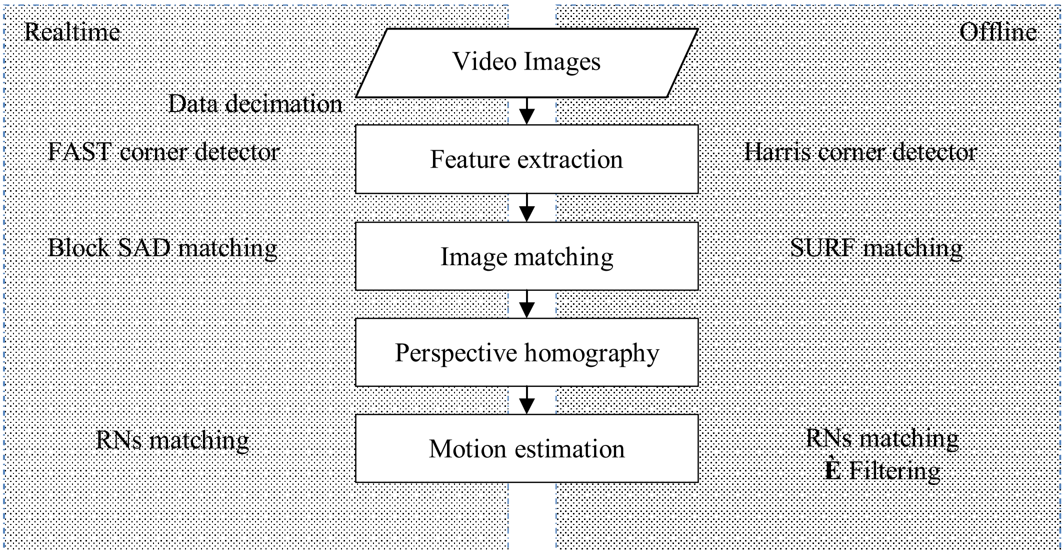

Figure 1: Working flow of proposed algorithm

2.1 Data decimation and feature extraction

The camera is carried by a person to capture a sequence of images looking directly at the ground while moving forward. The most common video frame frequency is about 30 frames per second, but it is not efficient and necessary to use all the recorded frames, especially for real-time processing. According to the references and testing experiences, decimating the data from 30 frames per second to 5 frames per second or even lower are suitable for pedestrian movement. The overlap rate between selected frames can remain 60% for matching. For real-time processing, the image resolution can be down sampling to increase the processing speed.

In this research, Features from Accelerated Segment Test (FAST) corner detector and Harris corner detector were used for feature points detection with different strategy. FAST corner detector can provide computational efficient results, but having anisotropic response and respond too readily to diagonal edges. As for Harris corner detector, the autocorrelation matrix (A) was calculated with Gaussian window to reduce the noise. And then, construct the cornerness map by calculating the cornerness measure (C) for each pixel from eq. (1), where k is a constant from 0.04 to 0.06 give the best results from empirical testing. With a threshold given and non-maximal suppression, the local maxima pixel can be extracted as a feature point.

$$C = \mathrm{Det}(A) - k * \mathrm{Trace}(A)^2 \qquad \text{Eq. (1)}$$

2.2 Image matching

After extracting feature points between subsequent images, Sum of Absolute Differences (SAD) was used to quickly find correspondences of the feature points between subsequent images. The formula was shown in eq. (2). It works by taking the absolute difference between each block-based pixels in the target window and the corresponding pixels in the searching window being used for comparison. These differences are summed to create a simple metric to check similarity.

$$E_{SAD}(u) = \sum_i |e_i| = \sum_i |I_1(x_i + u) - I_0(x_i)| \qquad \text{Eq. (2)}$$

For off-line processing, Speeded Up Robust Features (SURF) descriptor can provide more robust information for feature points between subsequent images. Advantages of SURF are scale and rotation invariant, resistance of noise in position and intensity, high match precision. The formula of each sub-region has a four-dimensional descriptor vector v was shown in eq. (3). For all 4 by 4 sub-regions can be concatenated as a descriptor vector of length 64. By matching this 64 dimensional descriptor, the conjugate points between subsequent images can be highly matched.

$$v = (\square\, d_x, \square\, d_y, \square\, |d_x|, \square\, |d_y|) \qquad \text{Eq. (3)}$$

2.3 Perspective homography

A perspective homography (H), which is a 3x3 matrix, can be used to mapping one image to another if the co-planar feature points between subsequent images are fully matched. The formula of this projective transformation is shown in eq. (4). H can be divided into two parts, metric part (M) and non-metric part (N). The degree of freedom are four in each. Non-metric part, N, is using for calibrating orthogonality and parallelism for the line segments with projective transformation parameters l1 and l2. Calibrating skewness and aspect ratio by affine transformation parameters α and β. Matrix (M) are composed with a 2x2 rotation matrix (R), a 2x1 translation vector (t), and an isotropic scaling (s). Camera motion vector can be found by t, orientation determined by θ as shown in eq.(5) and scaling factor eq.(6).

$$H = MN; \quad M = \begin{bmatrix} sR & t \\ 0^T & 1 \end{bmatrix}; \quad N = \begin{bmatrix} \frac{1}{\beta} & -\frac{\alpha}{\beta} & 0 \\ 0 & 1 & 0 \\ l_1 & l_2 & 1 \end{bmatrix} \qquad \text{Eq. (4)}$$

$$\theta = (\mathrm{atan}(R(1)/R(2)) + \mathrm{atan}(R(3)/R(4)))/2 \qquad \text{Eq. (5)}$$

$$s = ((R(1) + R(4))/\cos\theta)/2 \qquad \text{Eq. (6)}$$

However, features from the pedestrian's moving feet, shadows or noises should not used in the homography calculation. The RANSAC algorithm is used to remove the outliers. Random sample consensus using iterative calculation to estimate opitimal parameters of a mathematical model from observations which contains outliers. The procedure is listed as follow:

- Initialize transformation matrix
- Decide sampling times (E), inliers ratio or mount
- While count < E
    - Count = count + 1
    - Randomly select 4 pair of points from subsequent feature points
    - Calculate homography H, from the selected points
    - Replace H', if H has a distance metric less than H'
- Mapping all feature point pairs in both subsequent images by H to calculate a refined homography.

2.4 Motion estimation

In order to recognize the reference nodes as control points along the test path, this paper combines a pattern recognition method with identifications of reference nodes for calibrating the location estimation. As is well known, the simplest area-based image-matching method is the NCC algorithm. The NCC scheme is widely used in image-processing applications, and it is against the brightness difference between the image and template due to lighting condition, and the useful equations of the NCC approach are as follows.

$$\overline{G_T} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} G_T(x_i, y_i)}{n \cdot m} \quad , \quad \overline{G_S} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} G_S(x_i, y_i)}{n \cdot m} \qquad \text{Eq. (7)}$$

$$\sigma_T = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} (G_T(x_i, y_i) - \overline{G_T})^2}{n \cdot m - 1}} \quad , \quad \sigma_S = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} (G_S(x_i, y_i) - \overline{G_S})^2}{n \cdot m - 1}} \qquad \text{Eq. (8)}$$

$$\sigma_{TS} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} [(G_T(x_i, y_i) - \overline{G_T}) \cdot (G_S(x_i, y_i) - \overline{G_S})]}{n \cdot m - 1} \quad , \quad r = \frac{\sigma_{TS}}{\sigma_T \sigma_S} \qquad \text{Eq. (9)}$$

Where $G_T(x, y)$ and $G_S(x, y)$ are the image mask of the grayscale of target window and search window, respectively; $\overline{G_S}$ and $\overline{G_T}$ are the means of the grayscale of target window and search window, respectively; m and n are the numbers of rows and columns, respectively; σ_T and σ_S are the standard deviations of the image mask of the target window and search window, respectively. σ_{ST} is the value of cross correlation; r is the NCC value, and it can indicate that the most likely video frame or time passes through the landmarks.

As is well known, in Gaussian distribution, a KF tracking scheme is an optimal recursive data processing algorithm under the assumption of linear state function and the noise. Therefore, a KF algorithm can be used to correct the errors of the initial sensor measurement and to estimate the path of camera movement. There are two steps in applying KF tracking scheme, prediction and correction. Prediction functions were listed in eq.(10) and eq.(11), where x^- is state prediction; $\hat{x}$ is state correction; A is a parameter matrix; P^- is priori estimate error covariance; $\hat{P}$ is posteriori estimate error covariance; Q is process noise covariance.

state prediction :

$$x_k^- = A\hat{x}_{k-1} \qquad \text{Eq. (10)}$$

state prediction covariance:

$$P_k^- = A\hat{P}_{k-1}A^T + Q \qquad \text{Eq. (11)}$$

Correction functions are listed in eq.(12), eq.(13) and eq.(14), where K_k is Kalman gain; H is relate the state to the measurement; R is measurement noise covariance; z is measurement; I is a identity matrix. From eq.(12), if the measurement error is small and can be neglect, R will equal to zero and K_k is equal to H^{-1}, so $\hat{x}_k$ will equal to z_k. In this case, the state function will totally be contributed from measurement and neglect prediction values. After the correction with different Kalman gain, the measurement error will be reduced and generated optimal motion estimation.

$$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1} \qquad \text{Eq. (12)}$$

$$\hat{x}_k = x_k^- + K_k(z_k - Hx_k^-) \qquad \text{Eq. (13)}$$

$$\hat{P}_k = (I - K_k H)P_k^- \qquad \text{Eq. (14)}$$

EXPERIMENTAL RESULTS

The test area of this study is the roof of the Center for Space and Remote Sensing (CSRSR) in National Central University (NCU), Taiwan. 26 well-distributed reference nodes were set in this test area, and all the position is measured by ground-based measurement. The complete path is a closure loop start and end at node no.1, walking counter clockwise for 98 meters long as illustrated in Figure 2. A, B, C and D in Figure 2 are the corner points along the path. More then 4700 images were captured with over 90% overlap between two subsequent images. The image resolution is 480x640 pixels with fixed focal length. The video is recorded as Audio Video Interleaved (AVI) format. In order to decrease data amount, the frame span was set as 5 in real-time calculation. Figure 3 shows the overlap area with subsequent images between spanned frames.

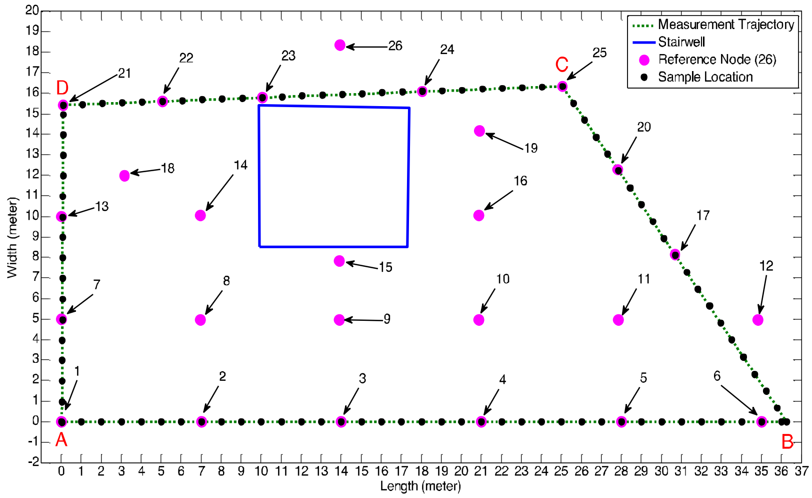

Figure 2: Reference nodes distribution and walking path

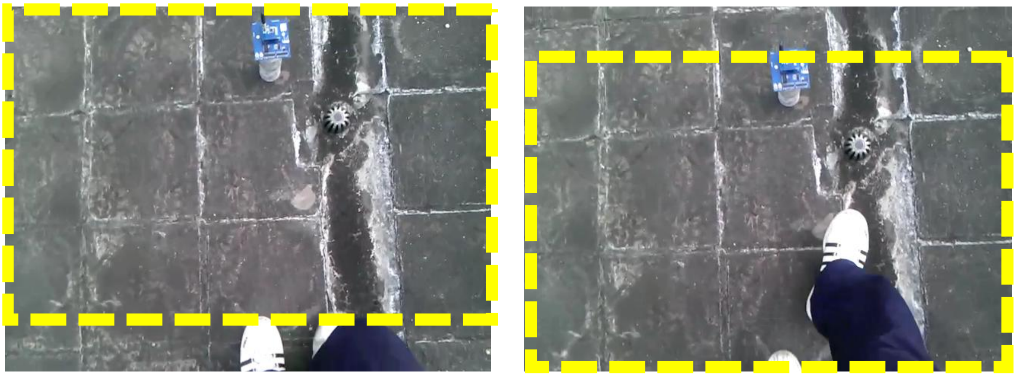

Figure 3: Overlap area with subsequent images between 5 spanned frames

Camera started from corner A walking straight to B, C and D counter-clockwisely and finally return to corner A. The translation between frame to frame was shown in Figure 4, started from top to the bottom of the figure. There are 3 "N" shape trajectories respectively represent three corner position when making a left turn. One barrier area between node 7 and node 1 in the nearly end of the path was obviously estimated when the carrier is avoiding the block. Rotation angles between frames was calculated and illustrated in Figure 5. Similar to the translation results, by accumulate the rotation angles, three corner position and barrier area can be easily detected and marked in Figure 6.

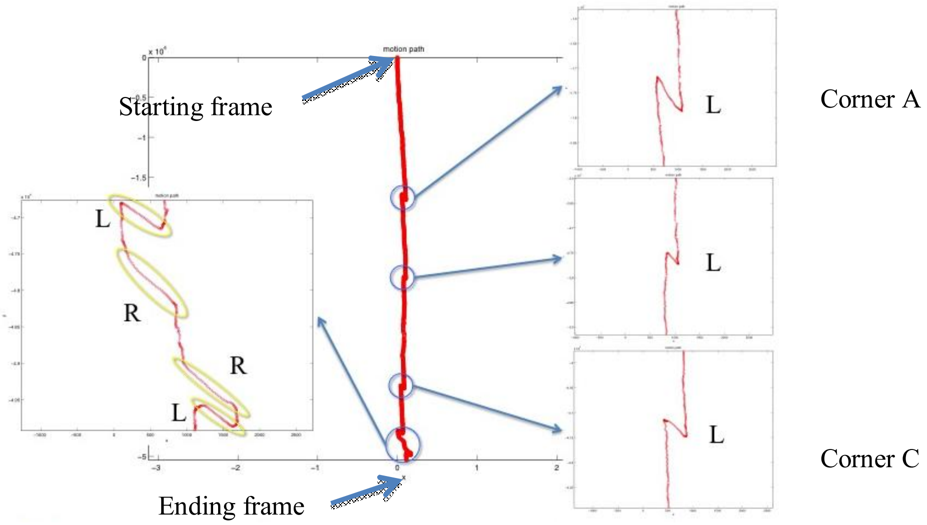

Figure 4: The translation estimation between frame to frame; Three corner position and barrier area was marked and enlarged to show the difference between left (L) and right (R) turns

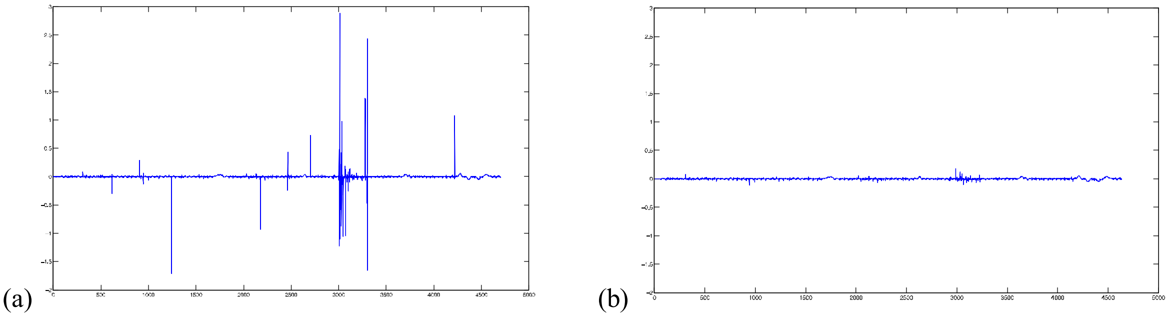

Figure 5: Rotation angle between frames (a) raw data, and (b) fileted unreasonable angle with given threshold; x axis is frame number, y axis is radians

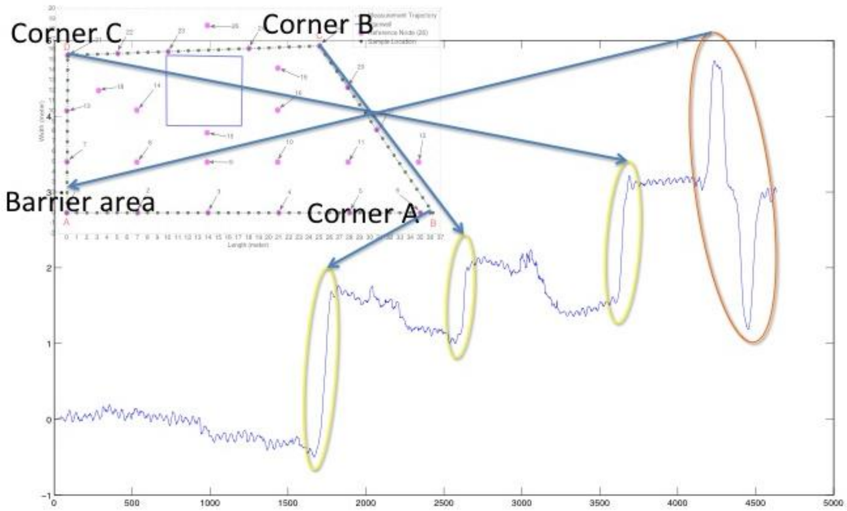

Figure 6: Rotation angle accumulation

The rotating angle at the corner position and barrier area can be measured according to the angle accumulation. Camera trajectory can be determined by Combing rotation and translation results as illustrated in Figure 7. This is the case if only vision-based method is used to estimate the camera trajectory. Obviously, there is heavy error propagation due to the rotation angle estimation. In order to reduce the drift effect, NCC was used to matching the template in Figure 8(a). When NCC value is local maximum and lager then threshold, the camera trajectory will force the current position to corresponding reference node. Figure 8(b) is the NCC value detected along the test path, and the matched reference node was marked as hollow circle.

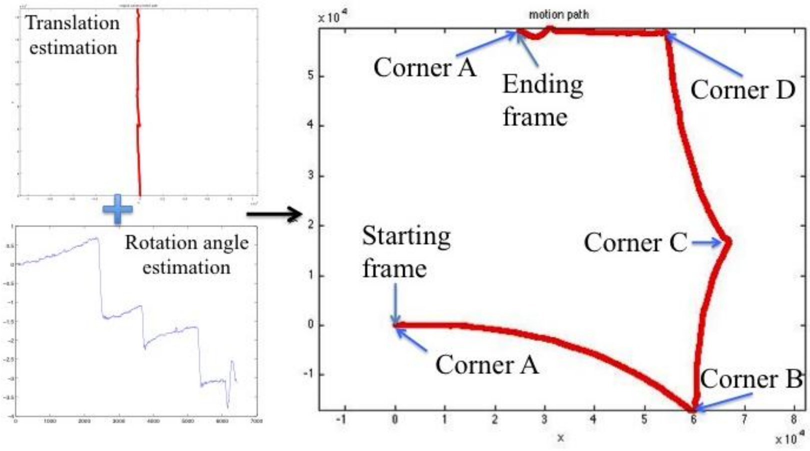

Figure 7: Motion estimation with vision-based method only

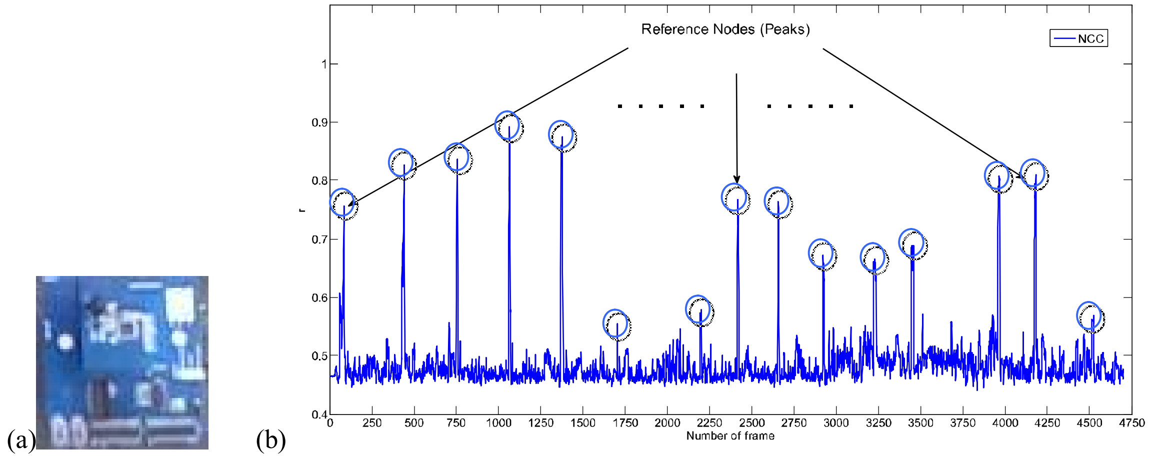

Figure 8: NCC matching to update trajectory position (a) matching template (b) NCC value detected along the test path

Red thick line in Figure 9(a) is the trajectory marked with NCC matched reference nodes. Compared to the ground truth drawn in blue thin line, the trajectory contains huge drift error due to the false angle propagation. NCC matched reference nodes can provide absolute coordinates to calibrate the drift effect. Figure 9(b) is the trajectory calibrated with referenced coordinates when camera pass through the reference nodes. Only a few drift effect remain because the camera still rotating after pass through the corner point. This is easily can be removed by applying a rotating filter. The average drift error along the path is less then 0.2 meters, and the maximum error is less then 0.9 meters.

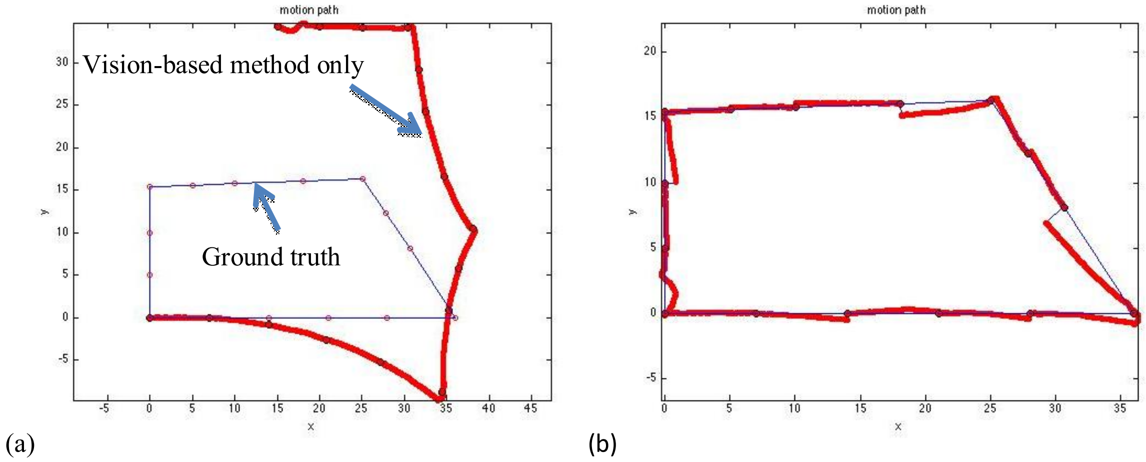

Figure 9: Trajectory calibrated with reference nodes (a) comparison between vision-based method only and ground truth data (b) trajectory calibrated with reference nodes

CONCLUSIONS

This paper presented a reduced-complexity location-estimation based on vision geometry. The advantage of vision-based tracking is that it can work in places where GPS is unavailable. This study combined the results of vision-based positioning and NCC matching to update the absolute coordinate. The proposed method can provide high frequency and reliable position of the camera trajectory only use continues video frames. The average drift error along the path is less then 0.2 meters, and the maximum error is less then 0.9 meters. The determined camera orientation and translation parameters can be used in various applications, such as providing initial conditions to real-time positioning and tracking in indoor or outdoor environments when GPS signal is lock-lose. Future work will focus on developing rotation angle filter to self-filtering small vibration and systematic drift. Also combine Microelectromechanical systems (MEMS) to obtain more information, in order to improve the positioning and tracking accuracy.

REFERENCES:

Bastanlar, Y. and Yardimci, Y., 2008. Corner Validation based on Extracted Corner Properties. Computer Vision and Image Understanding, pp. 243-261.

Bouguet, J.-Y., and Perona, P., 1995. Visual navigation using a single camera,“ Fifth International Conference on Computer Vision, pp. 645-652, 20-23 Jun 1995.

Chang, H., Lee, C.H. and F. Tsai, 2011, "Using Video Images to Obtain Camera Orientation and Translation Parameters", Proc. 32th Asian Conference on Remote Sensing (ACRS2011), Oct. 3-7, 2011, Taipei, Taiwan.

Hartley, R. and Zisserman, A., 2003. Multiple View Geometry in Computer Vision, second ed. Cambridge University Press, Cambridge, MA.

Davison, A. J., 2003. Real-time simultaneous localisation and mapping with a single camera. In Proceedings of the International Conference on Computer Vision (ICCV).

Strasdat, H., Montiel, J. M. M. and Davison, A. J., 2010. Scale drift-aware large scale monocular slam. In Proceedings of Robotics: Science and Systems.

Hide, C., Botterill, T., Andreotti, M., 2010. Low cost vision-aided IMU for pedestrian navigation. Ubiquitous Positioning Indoor Navigation and Location Based Service, pp. 1-7, 14-15 October 2010 Session 6.

Scaramuzza, D., Fraundorfer, F., Pollefeys, M. and Siegwart. R., 2009. Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints. In Proceedings of the International Conference on Computer Vision (ICCV).

Tsai, F., Chiou , Y.-S., Chang, H., 2012, " A Location-Tracking Testbed Using Vision-Assisted Scheme For Wireless Sensor Networks", Proc. XXII International Society For Photogrammetry & Remote Sensing Congress (ISPRS2012), Oct. 25- Sep.1, Melbourne, Australia.

ACKNOWLEDGEMENTS

This project was supported in part by National Science Council of Taiwan under grant Nos. NSC-99-2628-E-008-010. And partially supported by the CECI Engineering Consultants of Taiwan under project number 01935.