A feature-based approach for monocular camera tracking in unknown environments

HOSEINI, S.A. and KABIRI, P. <http://orcid.org/0000-0001-5143-0498>

A Feature-based Approach for Monocular Camera Tracking in Unknown Environments

Seyyed Ali Hoseini
School of Computer Engineering
Iran University of Science and Technology
Tehran, Iran
hoseiniali@iust.ac.ir

Peyman Kabiri
School of Computer Engineering
Iran University of Science and Technology
Tehran, Iran
peyman.kabiri@iust.ac.ir

***Abstract*—Camera tracking is an important issue in many computer vision and robotics applications, such as, augmented reality and Simultaneous Localization And Mapping (SLAM). In this paper, a feature-based technique for monocular camera tracking is proposed. The proposed approach is based on tracking a set of sparse features, which are successively tracked in a stream of video frames. In the developed system, camera initially views a chessboard with known cell size for few frames to be enabled to construct initial map of the environment. Thereafter, Camera pose estimation for each new incoming frame is carried out in a framework that is merely working with a set of visible natural landmarks. Estimation of 6-DOF camera pose parameters is performed using a particle filter. Moreover, recovering depth of newly detected landmarks, a linear triangulation method is used. The proposed method is applied on real world videos and positioning error of the camera pose is less than 3 cm in average that indicates effectiveness and accuracy of the proposed method.**

Keywords—Camera Tracking; Particle Filter; 3D reconstruction; visual SLAM.

I. INTRODUCTION

Purpose of the vision-based camera tracking is to estimate pose of the camera from a sequence of input images often in the form of video frames. Monocular Simultaneous Localization And Mapping (SLAM) and construction of the 3D representation of explored scene, have recently become popular field of research. Augmented reality and robot navigation are two major applications, which seriously utilize camera tracking and 3D reconstruction.

Unlike range scanners and RGB-D cameras, which produce favorable information about depth of observed scene, a monocular camera is a bearing-only sensor that only provides 2D measurements of a 3D environment. On the other hand, Euclidean estimation of camera pose, at least some information about depth of a sparse set of scene landmarks is necessary.

If loop closure is not exploited or no information in shape of known markers or fiducials are received from the scene, then the problem is considered as a dead reckoning technique. Thus, estimation of the camera pose and the 3D position of newly added features are calculated in accordance to previous estimations. This condition for long-range sequences may lead to uncontrolled accumulation of error.

In general, there are two main strategies to address the problem of camera tracking, feature-based methods and direct methods. In feature-based methods, the problem carries out in two steps. First, a limited number of features are extracted from images according to a saliency criterion. Second, camera pose and 3D geometry of the observed environment is estimated using the extracted features in previous step. Conversely, in direct methods, all the pixels within an image are exploited to estimate the camera position and orientation. Moreover, when the number of extracted features is low due to lack of texture, using direct methods provides more information about geometry of the environment.

A. Pose parameters

As depicted in Fig. 1, a moving camera captures images of environment from arbitrary positions. The main goal of the proposed system is to estimate position and orientation of the camera at each position with respect to the world referential system.

$$X_c = RX_w + t \tag{1}$$

X_c, X_w are the coordinates of the point with respect to camera and world coordinate systems, respectively. In the proposed system, it is aimed to estimate camera pose for successive frames of a captured video.

Structure of this paper is as follows: Related works are discussed in section II. In section III, the proposed approach

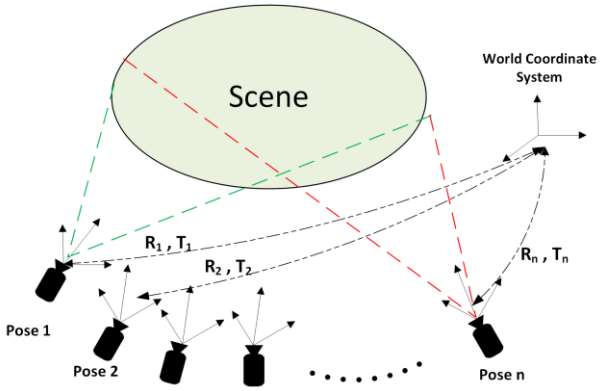

Fig. 1. Multi-view camera pose estimation.

will be explained in details. Experimental results are presented in section IV. Conclusions and future works are included in section V.

II. Related works

Camera tracking in unknown environments is a challenging task in computer vision and robotic research communities. Absence of markers or any pre-calibrated features in the scene produces cumulative error for camera pose parameters. Loop closure techniques [1, 2] may compensate for camera trajectory drift, but they are only desirable for situations where the camera revisits previously observed areas. It is necessary to detect and initialize new features once camera explores new regions for retrieving camera pose. In the reported work, detecting loops is not addressed. The reported work is focused on propagation of scene depth information to newly detected features as the camera observes new regions.

It is well known that receiving no information about depth of extracted scene features produces drift in camera trajectory and increases cumulative error. That is, for freely moving camera, captured images provide information about geometry of scene that can be recovered up to a scale factor using multi-view geometry. In this way, dealing with this problem, some researches put markers or fiducials with known structures in the scene to control cumulative error [3, 4]. Using multiple markers in the scene could also increase accuracy of camera pose parameters [5].

Exploiting reference-calibrated images is another technique for controlling growth of the camera pose error [6, 7]. Calibrated images are those with known 3D coordinates for a sparse set of features.

Generally, there are two main solutions to address the problem of camera tracking, i.e. Structure from Motion (SfM) and stochastic filtering. SfM approaches are mainly relied on techniques developed in multi-view geometry. In the core of these techniques, there is a group of algorithms appropriately explained on the basis of epipolar geometry [8]. Algorithms extended for camera pose estimation as well as techniques presented for 3D reconstruction were mostly applied on a small set of images. However, there are some reported works that are extended for longer image sequences [9, 10]. Moreover, they are often implemented in offline manner. Refining estimated parameters of the camera and 3D coordinates of the mapped features, often require additional optimization stage. Bundle Adjustment (BA) [11] and pose map [12] are two main strategies for this purpose.

In stochastic filtering approaches, problem is solved using the notion of dynamic systems. In this group of solutions the internal state of a dynamic system constitutes the parameters of camera motion. Furthermore, the state transition of the system is usually a linear relation based on physical nature of rigid body motion in 3D space and feature correspondences are utilized to form an observation model of the system. Mostly, due to nonlinear nature of the observation model, variants of Kalman filter such as Extended Kalman Filter (EKF) and Unscented Kalman Filter (UKF) are used for pose estimation [13, 14]. Particle Filter (PF) is another solution in the context of dynamic systems which is utilized for this purpose [15].

III. Proposed system

In Fig. 2 the overall scheme of the proposed system is illustrated. After arrival of new incoming frame, the process of camera pose estimation is performed in two stages, obtaining matched features and estimation of camera pose parameters. To provide robust matchings, extraction of salient and repetitive feature points is necessary. Feature detection and tracking are elaborated in subsequent sections. Thereafter, the extracted feature points should be matched with that of previous image. However, from previous image, only those features with known positions should be considered. Obtained matched pairs that are robust enough are used for estimation of camera pose parameters.

A. PF Implementation

In the proposed framework, PF is employed to estimate posteriori density for camera pose parameters. 6-DOF camera pose that consist of translation and rotation of camera with respect to world coordinate system, constitutes the state of PF and is denoted by $x_k = [t_k\ w_k]$, where t_k is the translation vector and the rotation matrix is encoded in w_k. w_k represents a rotation around the vector w_k with rotation angle equal to $|w_k|$. In the developed PF a constant position and orientation model is considered for state transition between time steps. Here it is assumed that camera pose only undertakes a Gaussian random walk with mean x_{k-1} and covariance matrix Σ_x as described in Eq. (1).

$$p(x_k \mid x_{k-1}) \propto N(x_{k-1}, \Sigma_x) \tag{1}$$

Let $Z_k = \{z_1, z_2, ..., z_n\}$ is a set of 3D points in the scene that are already initialized in the constructed map. Since camera is freely moving within a 3D space, in each frame, some feature points may go out of the camera's field of view and in return some new features are detected and further initialization is needed. This fact requires that Z_k to be updated

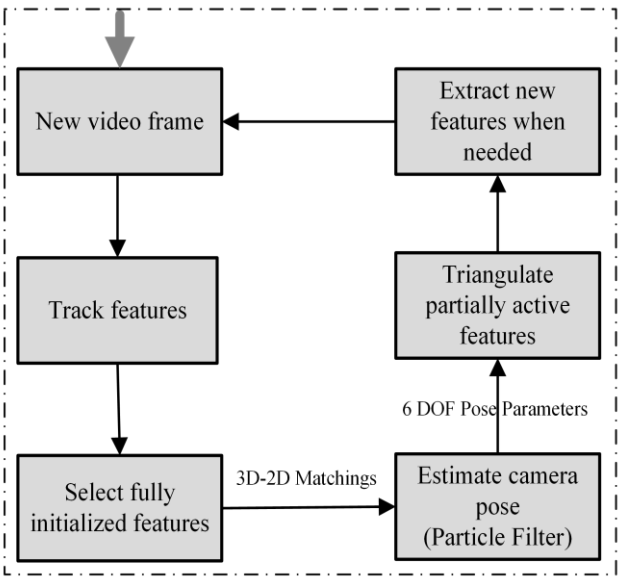

Fig. 2. Overview of proposed method.

continuously. It is also assumed that $U_k = \{u_1, u_2, ..., u_n\}$ is the associated observed feature points in frame k obtained through feature tracking procedure. Moreover, according to pinhole camera model for camera state x_k, the projection of $z_i \in Z_k$ on frame k is calculated using Eq. (2).

$$y_i^k = P_k \mathbf{z_i} = K\left(R_k z_i + t_k\right) \quad (2)$$

where $\mathbf{z_i}$ is homogenious representation of z_i and P_k is the camera projection matrix in frame k. Furthermore, K is the camera calibration matrix, R_k is the rotation matrix obtained from w_k using Rodrigues' formula [16] and t_k is the translation vector.

Implementing PF is aimed on successive estimation of the posteriori density $p(x_k | y_k, Z_k)$. This density is approximated as a weighted sum of samples drawn from state space as shown in Eq. (3).

$$p(x_k | y_k, Z_k) = \sum_{i=1}^{m} w_k^i \delta(x - x_k^i) \quad (3)$$

Here, m is the number of particles and x_k^i is the i-th sample drawn from state apace. w_k^i is weight of the x_k^i that is proportional to conditional likelihood $p(y_k|x_k,z_k)$. Given x_k^i, the i-th particle in frame k, w_k^i is computed by Eq. (4).

$$w_k^i = p(y_k | x_k^i, Z_k) \propto exp(-\lambda \sum_{i=1}^{n} (u_i - y_k^i)^T (u_i - y_k^i)) \quad (4)$$

Computed particle weights are scaled in such a way that $\sum_{i=1}^{m} w_k^i = 1$. Camera pose is then computed as the weighted sum of particles. It is evident that these weights constitute a probability distribution. In the next frame, the particles are sampled according to their weights through importance sampling.

B. System Initialization

In the reported work, a chessboard with known and equal cell sizes that is placed on a desk was used to estimate camera pose for initial few frames (about 10-15 frames). As depicted in Fig. 3, origin of the world coordinate system is aligned to one corner of the chessboard. Since size of the chessboard cells are known, they are selected as feature points with known 3D position in the world coordinate system. From these corner points and their projections on each frame, a collection of 3D-2D correspondences are supplied within each frame that makes it possible to calculate the camera pose with high precision. At the same time, extracted natural feature points detected in the first frame are tracked. Recovering depth information for the detected landmarks in the first frame completes the map initialization phase. In the reported work, only natural landmarks are used to retrieve camera pose parameters.

C. Feature extraction, tracking and initialization

In the proposed approach, FAST feature points [17] are detected and then tracked. Feature point tracking is a significant problem in camera tracking. Correlation window is the basic approach for tracking feature points in a sequence of consecutive video frames. However, in the reported work, a pyramidal technique for feature tracker is used [18]. Using this method to track each feature, a pyramidal representation for window with specified width centered at the associated feature is first constructed. Process of feature tracking is performed from the coarsest level to the finest one. At each level, the motion vector for each feature is calculated using the well-known Lucas-Kanade method for optical flow computation [19]. Result of the previous level is set as initial guess for the iterative registration of Lucas-Kanade method in the next level. Output of the aforementioned tracking method is a motion vector that represents displacement of the tracked feature. Number of pyramid levels is usually dependent on image resolution, however, for VGA quality images, 3 or 4 pyramid levels are convenient values.

An important property of the proposed system is the ability to add new natural landmarks to the map and then to estimate their 3D coordinates. Once a new feature is detected, it cannot be initialized immediately. This is due to the fact that, a single image does not hold any information about the depth of its points. In other words, it is necessary to track feature points along subsequent frames. Moreover, due to narrow-baseline nature of successive frames in a given video, linear triangulation of newly extracted features in two successive frames leads to remarkable error in depth calculation. To deal with this problem, a delayed initialization routine is employed. In other words, once a new feature is detected, its position is recorded. On subsequent frames, the feature is tracked and it is initialized once distance of its position on current frame from its position on the frame that was detected the first time, exceeds a predetermined threshold. In the reported experiments, this threshold set to 20 pixels.

IV. EXPERIMENTAL RESULTS

Experiments were carried out on two image sequences that are detailed in TABLE I. Sequences are captured at 30 frames per second rate. Observed scenery is a computer desk cluttered with various objects to generate images with rich textures.

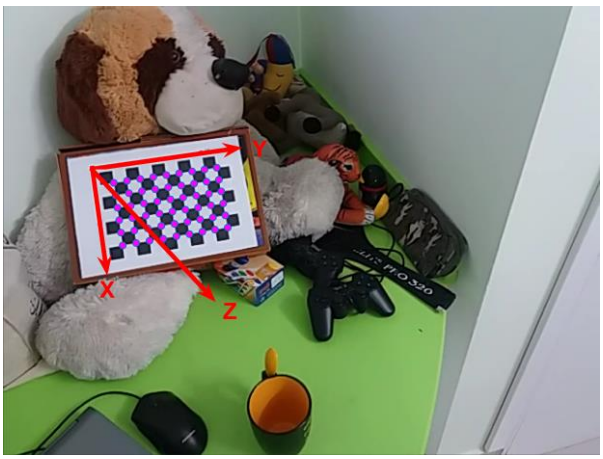

Fig. 3. Definition of world coordinate system and detected corners of the chessboard.

These rich textures enable the algorithm to extract required feature points. Furthermore, a chessboard pattern with known and equal cell sizes is embedded in the scene. The chessboard acts as a planar marker that its cells corners can be easily detected. Since the upper left corner of chessboard is considered as the origin of world coordinate system (Fig. 3), positions of the other cells corners are obtained effortlessly. Hence, by detecting projection of this points on captured images, a collection of 3D-2D feature correspondences is provided. Using these 3D-2D accurate feature correspondences, the Ground-Truth camera pose with high precision is calculated that enables us to evaluate performance of the proposed approach. Additionally, it is assumed that the camera is already calibrated. Camera calibration is performed using a technique presented by Zhengyou [20]. In TABLE II, intrinsic parameters of the camera are presented.

TABLE I. SPECIFICATON OF USED SEQUENCES

	Resolution	Number of Frames
Seq 1	640x480	1043
Seq 2	1280x72	1230

Fig. 4 shows the estimated camera trajectory against ground-truth in XY plane. As shown, in both sequences the camera center location is smoothly tracked along the camera path. In Fig. 5, components of the estimated camera position against ground-truth data are depicted. As it is shown, despite the long length of the input video sequences, camera moving path is tracked with high accuracy.

Furthermore, results produced by the reported approach are compared versus other tracking algorithms using camera pose estimations by EKF, UKF and EPnP [21] methods and the underlying results are reported in TABLE III. The presented results are in terms of Root Mean Square Error (RMSE) of translation and rotation components of camera pose along all sequences frames. One can see that the proposed PF-based approach outperforms the other nonlinear filtering approaches (EKF and UKF) as well as the EPnP method.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, a feature-based camera tracking approach in unknown environments is reported. A PF framework that operates on the basis of tracked FAST feature points is employed to estimate the 6-DOF state of the camera. Moreover, for better handling of camera quick motions, a pyramidal scheme for feature tracking system is used.

A significant property of the proposed system is that, it does not require any known marker in the scene while the algorithm is running. Of course, it should be kept in mind that when the number of video frames is increased the cumulative error for orientation and translation of camera will increase as well. This issue will introduce drift in camera trajectory, which directly affects triangulation accuracy. To overcome this problem, it is required to either acquire some information from the scene or try to close the loop. Intention is to consider the latter case as for the future work.

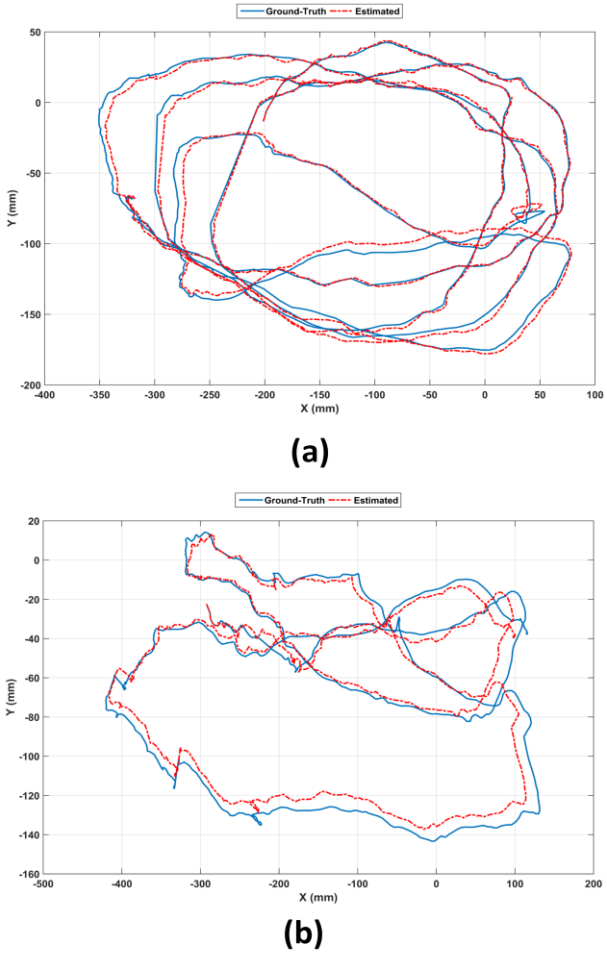

Fig. 4. Estimated and Ground-Truth camera trajectory on XY Plane for (a) Seq 1 (b) Seq 2.

TABLE II. INTRINSIC PARAMETERS OF CAMERA

Projection Parameters	
Scaling factors	***Principal point coordination***
$f_x = 517.89$ $f_y = 515.43$	$o_x = 321.65$ $o_y = 237.88$
Distortion parameters	
Radial distortion coefficients	***Tangential distortion coefficients***
$k_1 = 0.324$ $k_2 = -1.277$	$p_1 = 0$ $p_2 = 0$

TABLE III. RMSE OF TRANSLATION AND ORIENTATION

Pose Estimation Method	**Seq 1**		**Seq 2**	
	RMSE of Translation (mm)	***RMSE of Rotation (deg)***	***RMSE of Translation (mm)***	***RMSE of Rotation (deg)***
PF	16.7	3.5	25	6.4
EKF	197	7.2	243.4	20.5
UKF	92.8	8.3	291.2	24.3
EPnP	200.7	3.7	186.2	4.1

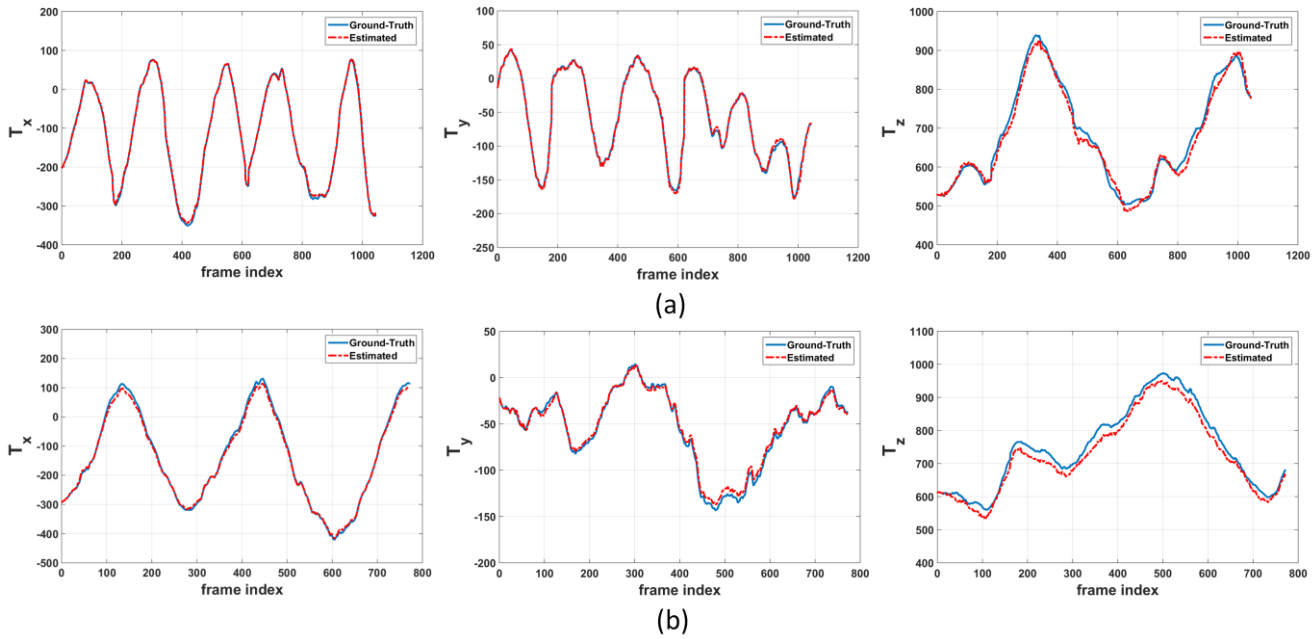

Fig. 5. Estimated camera pose against Ground-Truth data (a) Seq 1 (b) Seq 2.

REFERENCES

[1] L. A. Clemente, A. J. Davison, I. D. Reid, J. Neira, and J. D. Tardós, "Mapping Large Loops with a Single Hand-Held Camera," in *Proceedings of Robotics: Science and Systems*, 2007.

[2] E. Eade and T. Drummond, "Unified Loop Closing and Recovery for Real Time Monocular SLAM," in *BMVC*, 2008, pp. 136-145.

[3] F.-e. Ababsa and M. Mallem, "Robust camera pose estimation using 2d fiducials tracking for real-time augmented reality systems," presented at the Proceedings of the 2004 ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry, Singapore, 2004.

[4] M. Maidi, J.-Y. Didier, F. Ababsa, and M. Mallem, "A performance study for camera pose estimation using visual marker based tracking," *Machine Vision and Applications*, vol. 21, pp. 365-376, 2010.

[5] J.-H. Yoon, J.-S. Park, and C. Kim, "Increasing Camera Pose Estimation Accuracy Using Multiple Markers," in *Advances in Artificial Reality and Tele-Existence*. vol. 4282, 2006, pp. 239-248.

[6] K. Xu, K. W. Chia, and A. D. Cheok, "Real-time camera tracking for marker-less and unprepared augmented reality environments," *Image and Vision Computing*, vol. 26, pp. 673-689, 2008.

[7] Z. Dong, G. Zhang, J. Jia, and H. Bao, "Efficient keyframe-based real-time camera tracking," *Computer Vision and Image Understanding*, vol. 118, pp. 97-110, 2014.

[8] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*: Cambridge University Press, 2003.

[9] A. W. Fitzgibbon and A. Zisserman, "Automatic camera recovery for closed or open image sequences," in *Proceedings of 5th European Conference on Computer Vision*, 1998, pp. 311-326.

[10] M. Pollefeys, R. Koch, and L. Van Gool, "Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters," in *Sixth International Conference on Computer Vision*, 1998, pp. 90-95.

[11] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle Adjustment — A Modern Synthesis," presented at the International workshop on vision algorithms, 1999.

[12] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g 2 o: A general framework for graph optimization," in *IEEE International Conference on Robotics and Automation*, 2011, pp. 3607-3613.

[13] S. Jain and U. Neumann, "Real-time Camera Pose and Focal Length Estimation," in *18th International Conference onPattern Recognition*, 2006, pp. 551-555.

[14] M. Maidi, F. Ababsa, M. Mallem, and M. Preda, "Hybrid tracking system for robust fiducials registration in augmented reality," *Signal, Image and Video Processing*, vol. 9, pp. 831-849, 2015.

[15] J.-S. Kim and K.-S. Hong, "A recursive camera resectioning technique for off-line video-based augmented reality," *Pattern Recognition Letters*, vol. 28, pp. 842-853, 2007.

[16] R. M. Murray, S. S. Sastry, and L. Zexiang, *A Mathematical Introduction to Robotic Manipulation*: CRC Press, Inc., 1994.

[17] E. Rosten and T. Drummond, "Fusing points and lines for high performance tracking," in *Tenth IEEE International Conference on Computer Vision*, 2005, pp. 1508-1515.

[18] J.-Y. Bouguet, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm," 2001.

[19] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," presented at the Proceedings of the 7th international joint conference on Artificial intelligence, Vancouver, BC, Canada, 1981.

[20] Z. Zhengyou, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1330-1334, 2000.

[21] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An Accurate O(n) Solution to the PnP Problem," *International Journal of Computer Vision*, vol. 81, pp. 155-166, 2009.