Self-Adaptive Ground Calibration in Binocular Surveillance System

Diwen Liu*, Ling Cai† , Yuming Zhao* and Fuqiao Hu*
*School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
Email: {1130329107, fqhu, arola_zym}@sjtu.edu.cn
†School of Information Science and Engineering, Xiamen University
Email: cailing.cs@gmail.com

***Abstract*—Object detection and tracking have always been crucial and challenging topics in computer vision. Compared with monocular vision systems, binocular vision systems (BVSs) have the advantage of dealing with illumination variation, shadow interference, and severe occlusion. Usually, the BVS constructs the world coordinates system by manually calibrating the ground plane. However, the camera vibrations decreases the calibration precision and weakens the system performance. To automatically correct and update the parameters of ground plane, we introduce Linear Discriminant Analysis (LDA) method to analyze the results of object localization and include the feedback in the surveillance system, in this way, a close loop system that greatly improves the accuracy and stability of surveillance system is constructed. Experimental results demonstrate that our approach works well in BVS for video surveillance.**

***Index Terms*—Binocular vision, Multi-object tracking, Self-adaptive ground calibration, Linear Discriminant Analysis**

I. Introduction

Object detection and tracking are the fundamental challenges in computer vision and pattern recognition. Some state-of-the-art surveillance systems follow the detection and tracking framework, i.e., combining detection models with tracking algorithms together for video surveillance [1]. Nevertheless, the majority of existing works are hard to be applied in real video surveillance scenarios. Typically, three challenges have been widely recognized: (1) illumination variations; (2) shadow interference; (3) multiple objects occlusion. However, BVSs [2], [3] are more appropriate to handle these challenges than monocular vision systems. For the surveillance systems with stationary cameras, the world coordinate system can be built up by the calibrated ground plane, i.e., recovering the 3D world coordinates of objects. Unfortunately, the camera vibrations will cause the unexpected changes in the calibrated ground plane, which further weakens the detection and tracking performance. In this paper, we aim to adjust the parameters of ground plane automatically by those objects which have been localized in the surveillance system.

A. Related works

In many surveillance systems, background subtraction is widely used to detect moving objects in a stationary scene. Moreover, some learning-based methods are proposed to localize some specific objects. R. Xiao, et al. use the boosting chain method [4] for human face detection. H. Grabner proposes an AdaBoost feature selection framework to conduct on-line feature selection [5]. To overcome the drawback of drifting, they explored the continuum between the fixed tracker and online learning methods to obtain a classifier [6]. R. Liu et al. find an upper bound of boosting error in a co-training framework to guide the tracker construction [7]. Besides, regularized least squares classifier [8] and Structural Risk Minimization classifier [9] are proposed to get better classifying results. All above works of monocular vision systems have made great contributions for object detection and tracking, but they still cannot perfectly handle issues of objects occlusion, shadow interference and illumination.

To deal with the above mentioned challenges, the BVSs [10], [11] are widely used in video surveillance. In general, BVSs can be classified into two categories, wide and short baseline, by the distance of adjacent cameras. Currently, a variety of applications of surveillance scenes choose the wide baseline system without demands of calibration for the system parameters [12], [13]. To avoid the disturbance of noise, the wide baseline systems usually estimate the correspondence directly based on a sparse set of feature points in different views. While the short baseline systems usually construct the depth map by finding the correspondence points in different camera views [14], [15]. However, the computational process of the dense depth map is costly and sensitive to noise. L. Cai et al. [2] proposed a kernel-based algorithm to detect and track objects in the short baseline system with a greatly decreased computational cost.

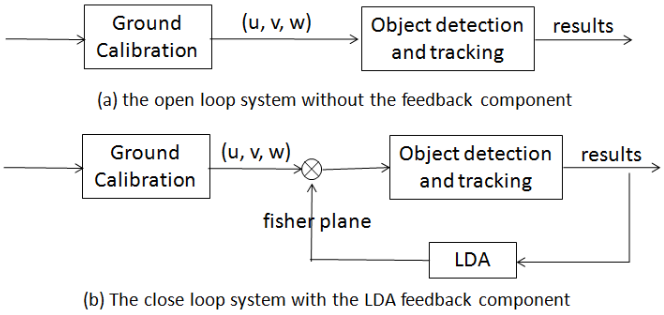
B. Our contribution

In stationary camera scenarios with a flat ground plane, L. Cai et al. [2] extract some sparse feature points for real-time object detection and tracking, but the system performance is closely related to the precision of the ground plane. Once the camera vibrates, the calibration result is not valid anymore, consequently the performance of system will be greatly affected.

To improve the accuracy and robustness of binocular surveillance system, we follow the assumption in [2], i.e, the surveillance system has the stationary cameras and the flat ground plane. But unlike the traditional open loop systems, we construct a closed loop surveillance system by adding a LDA-

† Corresponding author. This work is supported by NSFC 61572410.

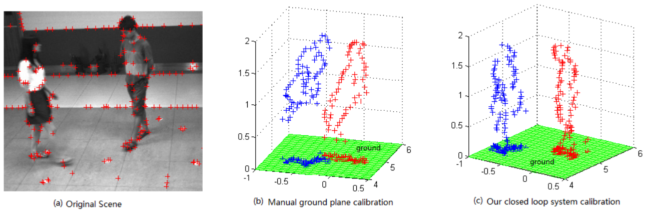
based [16] feedback component to achieve better performance, as shown in Fig.1. Given the feature points on the objects being identified in our surveillance system, we can project them onto the ground plane so that the distance of projected points of different objects are maximized and the distance of projected points of the same object is minimized. That is the ground plane can be regarded as the fisher plane in LDA algorithm. Therefore, we can utilize LDA algorithm to estimate the parameters of ground plane and update online to offset the camera vibrations, i.e., cameras shift or rotate within a small range, which is greatly beneficial for the robustness of the system. With the feedback of LDA, the projected points on the ground plane have less overlap and can be easily discriminated, as shown in Fig.2. The rest of the paper is organized as follows: Section 2 details our binocular coordinate system and the proposed method for auto-correction of the ground calibration, and followed by an introduction of the tracking system based on stationary binocular cameras in Section 3. Experimental results are shown in Section 4. Finally we conclude this paper in Section 5.



(a) the open loop system without the feedback component

(b) The close loop system with the LDA feedback component

Fig. 1: (a) the open loop system without the feedback component. (b) The close loop system with the LDA feedback component.



(a) Original Scene

(b) Manual ground plane calibration

(c) Our closed loop system calibration

Fig. 2: The effects of LDA feedback. (a) The original scene image. (b) The world coordinate system without the feedback of LDA. (c) The world coordinate system with the feedback of LDA.

II. Ground plane calibration

In the camera view, objects are often occluded by their neighbours, and hard to be separated clearly. However, there is no object occlusion in the top view. Therefore, we project the 3D feature points onto the ground plane to obtain the points' position in the top view.

A. Binocular vision system

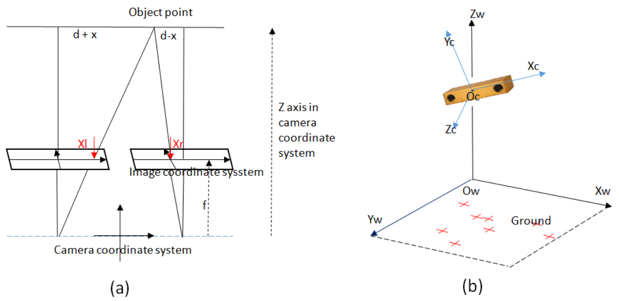


Fig. 3: Binocular vision system. (a) The relationship between the image coordinate system and the camera coordinate system, (b) The relationship between the camera coordinate system and the world coordinate system.

The overall binocular vision system is depicted in Fig.3. In general, the image coordinates are the projected position of physical object onto the image plane, while the camera coordinates are localizations of the physical object relative to the camera. Their relationship is shown in Fig.3 (a), and can be written as

$$\begin{cases} X_c = \frac{xZ_c}{f} = \frac{2dx}{x_l - x_r} \\ Y_c = \frac{yZ_c}{f} = \frac{2dy}{x_l - x_r} \\ Z_c = \frac{2df}{x_l - x_r} \end{cases} \tag{1}$$

where (X_c, Y_c, Z_c) denotes the camera coordinates, and (x_l, x_r) denotes image coordinates. f means the focal length of camera.

Fig.3 (b) shows the relationship between camera coordinates and world coordinates, it can be derived as:

$$(X_w, Y_w, Z_w)' = R(X_c, Y_c, Z_c)' + T \tag{2}$$

where R and T represents rotation matrix and displacement vector. To initialize R and T, we use those feature points on the ground plane (See Fig.3 (b) red crossing points) to fit the plane equation $uX_c + vY_c + wZ_c = 0$ by the least square method.

B. LDA feedback

The calibration error and cameras vibration will result in incorrect ground plane parameters and poor localization performance. For instance, Fig.4 shows the world coordinates of objects while the ground plane parameters contain small errors so that the feature points of different objects cannot be separated easily.

Due to the errors of ground plane parameters, the projected points of different objects on the calibrated ground plane will be partially overlapped. In fact, the ground plane is perpendicular to the objects on it, so the projected area of objects on the true ground plane is smaller than that on any

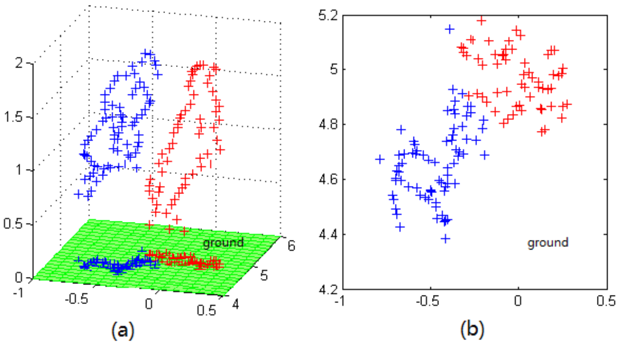


Fig. 4: Positions of feature points. (a) 3D position of feature points and calibrated ground plane. (b) Projected position of feature points on the calibrated ground plane.

other plane. Besides, the projected points of different objects on the true ground plane have no overlap, i.e., the scatter within an object of the same is minimal while among objects is maximal. Thus, we search the projected plane that separates the feature points belonging to different objects perfectly.

Assume that $C_1 = [x_1^1, x_1^2, ..., x_1^{n_1}]$, $C_2 = [x_2^1, x_2^2, ..., x_2^{n_2}]$, ...,$C_m = [x_m^1, x_m^2, ..., x_m^{n_m}]$, where, C_i stands for i-th objects, x_i^j means the j-th feature point belonging to the i-th objects. The targeted plane $\mathrm{W} = [\mathrm{w}_1, \mathrm{w}_2]^T$ maximizes the distance of projected sets $\mathrm{F}_1 = C_1\mathrm{W}^\mathrm{T} \in \mathbf{R}_{n_1\times 2}$, $\mathrm{F}_2 = C_2\mathrm{W}^\mathrm{T} \in \mathbf{R}_{n_2\times 2}$,..., $\mathrm{F}_\mathrm{m} = C_m\mathrm{W}^\mathrm{T} \in \mathbf{R}_{n_m\times 2}$. In order to determine this plane, we define the objective function as,

$$J(\mathrm{W}) = (\mathrm{W}\mathrm{S}_b\mathrm{W}^\mathrm{T})(\mathrm{W}\mathrm{S}_w\mathrm{W}^\mathrm{T})^{-1} \tag{3}$$

where S_w is the within-class scatter matrix and S_b is between-class scatter matrix.They can be defined as:

$$S_w = \sum_{i=1}^{m} \frac{n_i}{n} \sum_{j=1}^{n_i} (x_i^j - \bar{x}_i)(x_i^j - \bar{x}_i)^T \tag{4}$$

$$S_b = \sum_{i=1}^{m} n_i \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \tag{5}$$

where $n = \sum_{i=1}^{m} n_i$, $\bar{x}_i = 1/n_i \sum_{j=1}^{n_i}(x_i^j)$. LDA method is used to maximize Eq.3, i.e., constructing the two eigenvectors of $S_w^{-1}S_b$ with the two related biggest eigenvalues of W. To avoid the non full rank situation of S_w, we diagonalize it if necessary. The ground plane equation can be corrected by the fisher plane. Fig.2 (a) shows the corrected ground plane and the world feature points in the corresponding coordinate system. Fig.2 (b) shows the projected points on the ground plane of feature points.

III. Object Detection and Tracking

Most BVSs aim to construct the dense depth maps, but the high computational cost makes it hard to be applied in online video surveillance. Instead of recovering the dense depth map, we extract some harris [17] feature points to estimate their 3D position in the camera coordinate system.

With these 3D feature points, we project them onto the ground plane to generate the 2D projection points so that those 2D points can be grouped into different clusters to estimate the location and orientation of the objects. To model an object in the surveillance scene, we take the height of each point as weight for clustering. Then an object can be simply represented by an ellipse from the top view and the long axes of ellipses approximate its orientation.

A. Object detection

Kernel Density Estimation (KDE) is widely used to estimate the probability of the center point within a given window. Let the position and rotation of feature points be x and θ, and the probability density function can be described as

$$E(\mathrm{x}, \theta) = \sum_{i=1}^{n_i} H_\theta(d_i(\theta)) \sum_{j=1}^{n_j} w_j H_\mathrm{x}(d_j(\mathrm{x}), \theta_i) \tag{6}$$

where $d_i(\theta)$ and $d_j(\mathrm{x})$ are the distances of orientation and position with normalized coefficients. H_θ and H_x are the kernel functions of position and orientation respectively, i.e.

$$H_X\,(x) = 1 - x, (0 < x < 1) \tag{7}$$

$$H_\theta\,(x) = e^{-x} \tag{8}$$

A local maximum of function $E(\mathrm{x}, \theta)$ stands for the probability of position and rotation of an object. Mean shift method, proposed by Y. Cheng [18], is used to search for the local maximums. It is a hill climbing process along with the opposite direction of gradient in $E(\mathrm{x}, \theta)$. To search the local maximums of $E(\mathrm{x}, \theta)$, we iteratively update the orientation and position variables by

$$\widehat{\mathrm{x}} = \mathrm{x} + m_{h,G}(\mathrm{x}) \tag{9}$$

$$\widehat{\theta} = \theta + m_{h,G}(\theta) \tag{10}$$

The convergence point of mean shift iteration is the local maximums in $E(\mathrm{x}, \theta)$. Those points that climbs to the same peak are grouped into a cluster, i.e., to stand for the object. Fig.5 shows the mean shift iteration process and the cluster results.

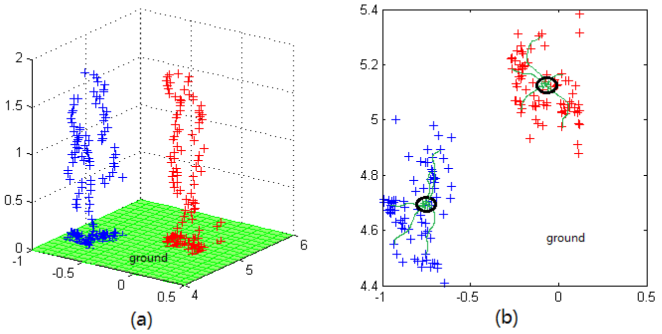
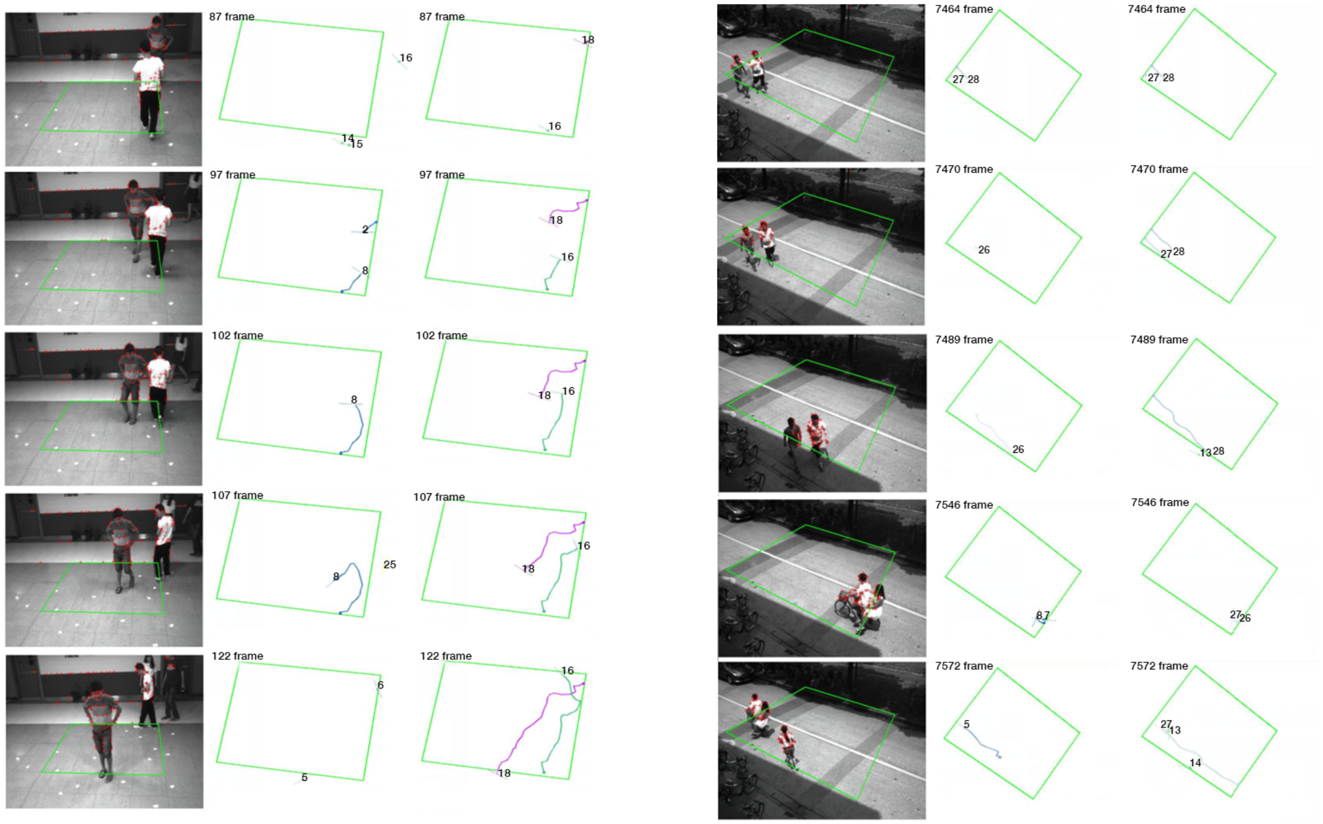


Fig. 5: 3D position of objects. (a) shows the 3D position of feature points and ground plane, (b) shows the iteration process and cluster results.



(a)Ground calibration with deviation and LDA auto-correction

(b)Cameras turbulence and LDA auto-correction

Fig. 6: trajectories of objects without LDA method and with LDA method in different scenarios

B. *Object tracking*

After localizing the objects in the scene, the object in consecutive frames need to be associated to generate its trajectory. Usually, two methods are used to track an object: detecting the object position in each frame by clustering algorithm and then associating them across frames; Or estimating the object position in a new frame by iteratively updating the position from the previous frame.

In this paper, we choose the second approach, i.e., employing Kalman filter to estimate objects' positions in consecutive frames by iteratively updating the position. The object position can be estimated by

$$V_t^{'} = V_{t-1} = x_{t-1}^{'} - x_{t-2}^{'} \tag{11}$$

where V_{t-1} is the velocity of objects at time $t-1$, $V_t^{'}$ is the predicted velocity of objects at time t, $x_{t-1}^{'}$ and $x_{t-2}^{'}$ are positions of an object at time $t-1$ and $t-2$. We can predict the position of the object at time t to initialize the hill climbing position.

The initial position of hill climbing in the current frame is estimated according to the filter prediction. Then the local maximum in the current frame is obtained by the mean shift iteration. At last, the position and orientation of the object is updated with the new local maximum. In this way, the trajectory of objects are obtained.

IV. Experiments

Both the indoor and outdoor experiments are conducted to demonstrate the effectiveness of employing LDA in our binocular vision tracking system. Fig.6 (a) presents the indoor experiments. The first column displays real scenario images, and the second column shows the tracking trajectories with large deviation of ground plane parameters. It can be seen that the tracking result is not accurate enough. While the third column shows the good tracking results when LDA is used to correct the ground plane automatically.

Fig.6 (b) shows the outdoor experiments. The first column displays the real scenario images and it can be seen that the field of view has been changed with camera vibrations. The second column shows the poor tracking performance without LDA correcting the ground plane equation automatically. However, the third column shows the robustness of tracking performance by LDA method.

Table 1 lists the experimental results under different scenes. Tracking results with LDA are compared to these without LDA method. It can be seen that LDA method is clearly beneficial to the robustness of the tracking system, and greatly improves the tracking performance.

TABLE I: tracking accuracy

Method	Scene 1	Scene 2	Scene 3	Scene 4
Cluster	**96.3%**	83.6%	79.5%	76.2%
Cluster+LDA	95.8%	**88.7%**	**87.2%**	**79.3%**

V. Conclusion

In this paper, we empoly LDA method in binocular vision tracking system to adjust the ground plane parameters automatically, and it can greatly improve the accuracy and robustness of BVS. The experimental results demonstrate the effectiveness of the proposed method.

References

[1] Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2012, 34(7): 1409-1422.

[2] Cai L, He L, Xu Y, et al. Multi-object detection and tracking by stereo vision[J]. Pattern Recognition, 2010, 43(12): 4028-4041.

[3] Tilneac M, Dolga V, Grigorescu S, et al. 3D Stereo Vision Measurements for Weed-Crop Discrimination[J]. Elektronika ir Elektrotechnika, 2012, 123(7): 9-12.

[4] Xiao R, Zhu L, Zhang H J. Boosting chain learning for object detection[C]//Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. IEEE, 2003: 709-715.

[5] Grabner H, Bischof H. On-line boosting and vision[C]//Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. IEEE, 2006, 1: 260-267. MLA

[6] Grabner H, Leistner C, Bischof H. Semi-supervised on-line boosting for robust tracking[M]//Computer VisionCECCV 2008. Springer Berlin Heidelberg, 2008: 234-247.

[7] Liu R, Cheng J, Lu H. A robust boosting tracker with minimum error bound in a co-training framework[C]//ICCV. 2009: 1459-1466.

[8] Rifkin R, Yeo G, Poggio T. Regularized least-squares classification[J]. Nato Science Series Sub Series III Computer and Systems Sciences, 2003, 190: 131-154.

[9] Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[M]//Computer VisionCECCV 2012. Springer Berlin Heidelberg, 2012: 702-715.

[10] Gavrila D M, Davis L S. 3-D model-based tracking of humans in action: a multi-view approach[C]//Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on. IEEE, 1996: 73-80.

[11] Pei Z, Zhang Y, Yang T, et al. A novel multi-object detection method in complex scene using synthetic aperture imaging[J]. Pattern Recognition, 2012, 45(4): 1637-1658.

[12] Mittal A, Davis L S. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene[J]. International Journal of Computer Vision, 2003, 51(3): 189-203.

[13] Khan S M, Shah M. A multiview approach to tracking people in crowded scenes using a planar homography constraint[M]//Computer VisionCECCV 2006. Springer Berlin Heidelberg, 2006: 133-146.

[14] Darrell T, Gordon G, Harville M, et al. Integrated person tracking using stereo, color, and pattern detection[C]//Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on. IEEE, 1998: 601-608.

[15] Huang X, Li L, Sim T. Stereo-based human head detection from crowd scenes[C]//Image Processing, 2004. ICIP'04. 2004 International Conference on. IEEE, 2004, 2: 1353-1356.

[16] Moreno-Noguer F, Sanfeliu A, Samaras D. A target dependent colorspace for robust tracking[C]//Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. IEEE, 2006, 3: 43-46.

[17] Harris C, Stephens M, "A combined corner and edge detector," [C]//Alvey vision conference. 1988, 15: 50.

[18] Cheng Y. Mean shift, mode seeking, and clustering[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1995, 17(8): 790-799.