

TRACKING OBJECTS FROM MULTIPLE STATIONARY AND MOVING CAMERAS

Jinman Kang, Isaac Cohen and Gerard Medioni

IRIS, Computer Vision Group, University of Southern California, USA
{jinmanka|icohen|medioni}@iris.usc.edu

Abstract

This paper presents a novel approach for multi-views tracking of moving objects observed by multiple, stationary or moving cameras. Video streams from stationary cameras are registered using ground plane homography obtained from known 3D ground plane information. In the more general case of heterogeneous cameras (a combination of stationary and Pan-Tilt-Zoom cameras), video streams are registered using a ground plane homography and affine transformations compensating camera motion. The detection of moving objects is performed by defining an adaptive background model that takes into account the camera motion approximated by the affine transformation. We address the tracking problem by modeling motion and appearance of the moving objects using probabilistic models. The object's appearance is represented using multiple colors distribution model that provides an efficient description of the object invariant to 2D rigid and scaling deformations. The motion models are obtained using a Kalman Filter (KF) process that predicts the position of the moving object in 2D, as well as in 3D when the images are registered to the ground plane. The tracking is performed by the maximization of a joint probability model reflecting object's motion and appearance. The novelty of our approach consists in modeling multiple trajectories observed by the moving and stationary cameras in the same KF framework, and integrating multiple cues and camera views in a Joint Probability Data Association Filter (JPDAF). The proposed approach allows deriving an accurate tracking of moving objects, an automatic camera handoff and the efficient management of partial and total occlusions. We demonstrate the performances of the system on several video sequences.

1 Introduction

Video surveillance of a large facility such as a warehouse, a campus or an airport usually requires a set of cameras for ensuring a complete coverage of the scene. Understanding human activity in such scenes requires the integration of video cues acquired by a set of cameras, as the activity usually unfolds in a large area that cannot be monitored by a single camera. Multiple stationary cameras are commonly used for monitoring large areas, as they provide wide coverage, good image resolution, and allow the inference of additional characteristics necessary for activity description and recognition. Omnicameras provide an alternative, but can only be used for close range objects in order to guarantee

sufficient image resolution for moving objects. Moving cameras such as pan-tilt zoom and hand-held cameras allow monitoring of dynamically changing regions of interest (ROI). They dynamically change the focus of the monitored region and provide a larger coverage than stationary cameras. Several algorithms for tracking moving objects across multiple stationary cameras have been proposed. Most of them use "color distribution" as the main cue for tracking objects across views [9][8][2]. Since color information can be easily biased by several factors such as illumination, shadow, blobs segmentation, appearance change, or different camera controls, color cue may be not very reliable for tracking moving objects across large scenes. In [3], the author proposed an approach for space and time self-calibration of cameras, but the proposed approach is limited to small moving objects and to top down views where the observed shapes are similar across views, occlusions are very rare, and the depth of the object is not significant. Multiple views tracking of moving objects from unsynchronized video streams was proposed in [5]. The authors proposed a spatio-temporal homography for registering unsynchronized views acquired from multiple stationary cameras. Automatic camera hand off control has been proposed in [10] using these principles.

We propose a novel approach for integrating information from multiple stationary or multiple heterogeneous cameras. We use a homography for registering stationary cameras on top of a known ground plane. This approach continuously tracks moving objects across two or more uncalibrated cameras. In the case of combining moving and stationary cameras, which is proposed in [6], an affine transform between a pair of frames is used for stabilizing moving camera sequences, and a perspective transformation for registering the moving and stationary cameras.

The detection of moving objects in the video stream is performed by a mode-based background method. Each detected moving blob is projected onto the ground plane for guaranteeing consistency of the labels across the cameras view points. The detection of moving objects in a video stream acquired by a moving camera is performed by a modified background learning method relying on a robust affine registration and a sliding window approach.

The tracking of detected moving objects in each view is formulated as a maximization of a joint probability model. The joint probability model consists of an appearance model, motion models. The appearance probability model is a color similarity measure between detected blobs. The appearance model is a color distribution model derived by segmenting the blob into a collection of polar bins. The appearance model is

defined as a combination of the polar distributions guaranteeing invariance to translation, rotation and scale of the object. The motion probability models are inferred from a Kalman Filter (KF). These models are calculated by a Gaussian distribution between the predicted bounding box position and the bounding box position of the observed blobs in 2D. For 3D, where a prior knowledge is available, a projection on the ground of the detected bounding box characterizes the observed objects' foot position. The predicted position of the foot on the ground plane is estimated using a KF.

The novelty of our approach consists in handling multiple cluttered and overlapped trajectories observed by multiple stationary or multiple heterogeneous cameras using JPDAF. It derives a simultaneous motion measurement for objects viewed by the two cameras, and allows for an automatic handling of occlusions, detection errors and cameras handoff. The paper is organized as follows. In Section 2 we describe the method for detection moving objects in video sequences acquired by a non-stationary camera. Section 3 describes the method for registering multiple cameras. Section 4 introduces the appearance-based model and the joint probability model used for tracking moving objects. The usage of the joint probability model for tracking is explained in Section 5. Obtained results are presented and discussed in Section 6. Finally, Section 7 concludes our paper with a discussion on future work.

2 Detecting moving regions

The main difference, between the detection of moving objects from a stationary and moving camera, is the characterization of the background model. In a stationary camera, variations in the image sequence are modeled at the pixel level and allow defining a background model for each pixel using statistical-based techniques. This concept can be extended to non-stationary cameras by compensating for the camera motion prior to the estimation of the background model. Registering the current frame to the selected reference is performed by concatenating the estimated pair-wise transforms.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix} \Rightarrow \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} a & b & t_x \\ c & d & t_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \quad (1)$$

$$\Rightarrow X_{t+1} = A_{(t,t+1)} X_t$$

where, $A_{(t,t+1)}$ is the affine transform from t to $t+1$.

Inaccuracies in the parameter estimation and the lack of proper approximation of the scene geometry by the affine model will inherently create registration errors and therefore erroneous background models. We propose using a sliding window where the number of frames considered is such as the sub-pixel accuracy is guaranteed by the affine registration. Moreover, for each current frame we re-warp the images in the buffer using the pair-wise registration and consequently minimizing the impact of an erroneous registration. Indeed an erroneous registration will not influence the quality of the detection for the whole sequence but only within the number of frames considered in the sliding window. The corresponding position of each pixel in the reference frame is re-estimated by:

$$\begin{aligned} cur > ref : X_{cur} &= A_{(ref, cur)} X_{ref} \\ &= A_{(cur-1, cur)} \cdots \cdot A_{(ref, ref+1)} X_{ref} \\ cur < ref : X_{cur} &= A_{(ref, cur)}^{-1} X_{ref} \\ &= A_{(cur, cur+1)}^{-1} \cdots \cdot A_{(ref-1, ref)}^{-1} X_{ref} \end{aligned} \quad (2)$$

In Figure 1, we show the detection results obtained with the proposed approach.



Figure 1. Comparison between two detection algorithms (a) Using a concatenation of affine transformation from the reference frame. (b) Using the proposed sliding window method

3 Registration of multiple cameras

Multiple stationary cameras provide a good coverage of the scene but require the integration of information across cameras. The geometric registration of cameras viewpoint is performed using a homography from a set of 4 matching points obtained from a ground plane model (e.g. a standard of soccer field). The main advantage of the proposed registration method is that we can not only relate detected moving regions in 2D, but also estimate the 3D position of the moving people. By combining the affine and perspective transforms, moving and static cameras can be spatially registered by a concatenation of homographies as depicted in Section 2.

4 Joint probability models

We decouple the tracking problem into two parts by first modeling objects' appearances by defining a color-based object representation and then modeling 2D and 3D velocities of objects. Each model is formulated as a probabilistic model, and the tracking problem is defined as maximization of the joint probability. A Joint probability data association filter (JPDAF) is used for combining the multiple hypotheses/cues extracted from the moving objects [11]. An appearance-based description of moving objects is used for measuring similarity among detected moving objects, while a KF is used for modeling 2D and 3D velocities of moving people.

4.1 Appearance model

Various methods have been proposed to solve the tracking problem using color and shape information. Commonly, a single color model we propose to use multiple color models and their relative localization for measuring efficiently the similarity between detected objects. In [1], a multiple color model was proposed for people tracking. It is based on subdividing the detected blob in three regions corresponding to the head, torso, and legs. The segmentation of the detected blob into body parts is ad-hoc and does not generalize to real life situations.

We propose an appearance model that is invariant to 2D rigid transformation. This model, defined by polar distribution, provides a description of object's colors properties. This 2D distribution model is used for measuring the similarity of tracked objects within and across cameras. The color distribution model is obtained by mapping the blob into a polar representation. Several shape or color distribution models using a polar representation have been proposed [4][6][7]. In [4], the proposed approach is focused on the object's shape description (edge) instead of their appearance (color), and it is only limited to representing local shape properties. In [6] the proposed model measures color distribution using a similar polar representation, but focuses on characterizing a global appearance signature of the object. The model is not 2D rotation-invariant and we propose here to use the shape description model proposed in [7] for guaranteeing invariance to 2D rigid transformation.

An illustration of the definition of the appearance model is shown in Figure 2 where we sampled the reference circle with 8 control points. The defined model is translation invariant. Rotation invariance is obtained by taking a larger number of control points along the reference circle. Finally, normalizing the reference circle to a unit circle guarantees scale invariance.

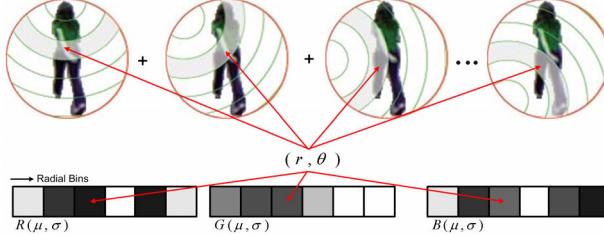


Figure 2. Illustration of the appearance model computation.

The appearance model probability is obtained by the average cross-correlation of color components between the appearance model of the tracked object and current observation.

4.2 Motion model

Motion models are approximated using a first order KF filter. We use the KF formalism to derive only the probability models associated to a constant velocity motion in the image plane (2D) and ground plane (3D). This probability will be then used by a joint probability model.

The moving object moves in the image space at constant velocity, and the new position is obtained by the following equation:

$$x_{t+1} = F_t x_t + e_t \Rightarrow \begin{bmatrix} x_T^{t+1} \\ y_T^{t+1} \\ u_T^{t+1} \\ v_T^{t+1} \\ x_B^{t+1} \\ y_B^{t+1} \\ u_B^{t+1} \\ v_B^{t+1} \\ x_{3D}^{t+1} \\ y_{3D}^{t+1} \\ u_{3D}^{t+1} \\ v_{3D}^{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_T^t \\ y_T^t \\ u_T^t \\ v_T^t \\ x_B^t \\ y_B^t \\ u_B^t \\ v_B^t \\ x_{3D}^t \\ y_{3D}^t \\ u_{3D}^t \\ v_{3D}^t \end{bmatrix} + \begin{bmatrix} e_1^t \\ e_2^t \\ e_3^t \\ e_4^t \\ e_5^t \\ e_6^t \\ e_7^t \\ e_8^t \\ e_9^t \\ e_{10}^t \\ e_{11}^t \\ e_{12}^t \end{bmatrix} \quad (3)$$

where, F is the system evolution matrix, e_t is the processing noise vector and x_t is the state vector considered in tracking moving object in 2D and 3D space. (x_T^t, y_T^t) and (x_B^t, y_B^t) are respectively the top-left and the bottom-right corner of the detected bounding box of the moving region and (u_T^t, v_T^t) and (u_B^t, v_B^t) are the corresponding 2D image velocity. We have discarded the z coordinates as we assume that this 3D point lies on the 3D ground plane used as origin. The coordinates (x_{3D}^t, y_{3D}^t) are computed from the detected moving blob: It corresponds to the centroid of blob's lowest pixels (i.e. of minimum y) and (u_{3D}^t, v_{3D}^t) corresponds to the velocity of the point (x_{3D}^t, y_{3D}^t) on the 3D ground plane.

The state vector considered in tracking moving object across heterogeneous cameras, is defined by the following vector:

$$x_i = (x_m^i, y_m^i, v_x^i, v_y^i, x_s^i, y_s^i, u_x^i, u_y^i) \quad (4)$$

where (x_m^i, y_m^i) is the lowest corners of the detected bounding box in the moving camera, (x_s^i, y_s^i) is the corresponding corner of the detected bounding box in static camera, (v_x^i, v_y^i) is the 2D velocity of (x_m^i, y_m^i) , and (u_x^i, u_y^i) is the 2D velocity of (x_s^i, y_s^i) .

By encoding camera motion into KF framework, we can adapt it for multiple heterogeneous cameras. The new position of the moving object across the heterogeneous camera set is obtained by the following equation:

$$x_{t+1} = F_t x_t + G_t + w_t \Rightarrow \begin{bmatrix} x_m^{t+1} \\ y_m^{t+1} \\ v_x^{t+1} \\ v_y^{t+1} \\ x_s^{t+1} \\ y_s^{t+1} \\ u_x^{t+1} \\ u_y^{t+1} \end{bmatrix} = \begin{bmatrix} a_t & b_t & a_t & b_t & 0 & 0 & 0 & 0 \\ c_t & d_t & c_t & d_t & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_m^t \\ y_m^t \\ v_x^t \\ v_y^t \\ x_s^t \\ y_s^t \\ u_x^t \\ u_y^t \end{bmatrix} + \begin{bmatrix} g_x^t \\ g_y^t \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} w_1^t \\ w_2^t \\ w_3^t \\ w_4^t \\ w_5^t \\ w_6^t \\ w_7^t \\ w_8^t \end{bmatrix} \quad (5)$$

where, F_t is the system evolution matrix at time t , G_t is the translation vector of the affine transform at time t , $a_t, b_t, c_t, d_t, g_x^t, g_y^t$ are the affine parameters of the affine transform at time t , and w_t is the processing noise vector.

The measurement equation only includes observed image position (e.g. corresponding bounding box position).

The 2D motion probability model (P_{2d_motion}), the 3D motion probability model (P_{3d_motion}) and the combinatorial motion probability model (P_{hetero}) are calculated by the Gaussian (normal) distribution of the motion estimates provided by the KF.

5 JPDAF-based tracking approach

We formulate the tracking problem as finding an optimized position in 2D (I') and 3D (W') of the moving object by maximizing all probability models. The joint probability P_{total} is defined by the product of the appearance and two motion probabilities

$$(I_0^*, W_0^*) = \arg \max_{(I_0, W_0)} P(C_0, I_0, W_0) \quad (6)$$

where C_0 denotes color observation, I_0 is the 2D position on the image space, and W_0 is the 3D position on the ground plane at time 0.

The optimal position at each time step depends on the current observation, as well as on the motion estimation at the previous positions. It corresponds to the maximum probability of all current and previous observations. Consequently, a joint probability of the given position at time t is formulated as follows:

$$P_{total}(I_t, W_t) = P(C_t, I_t, W_t, \dots, C_1, I_1^*, W_1^*, C_0, I_0^*, W_0^*) \quad (7)$$

Equation (7) can be decoupled into several components, and we obtained the joint probability of the current position by following equation:

$$\begin{aligned} & P(C_t, I_t, W_t, \dots, C_0, I_0^*, W_0^*) \\ &= P(C_t | I_t, W_t) P(I_t, | I_{t-1}^*, \dots, I_0^*) P(W_t | W_{t-1}^*, \dots, W_0^*) P_{total}(I_{t-1}^*, W_{t-1}^*) \\ &= P_{color}(I_t, W_t) P_{2D_motion}(I_t) P_{3D_motion}(W_t) P_{total}(I_{t-1}^*, W_{t-1}^*) \\ &= P_{total}(I_t, W_t) \end{aligned} \quad (8)$$

In the case of multiple heterogeneous cameras, the calculation of the joint probability can be simplified as follows:

$$\begin{aligned} & P(C_t, I_t, \dots, C_0, I_0^*) \\ &= P(C_t | I_t) P(I_t, | I_{t-1}^*, \dots, I_0^*) P_{total}(I_{t-1}^*) = P_{color}(I_t) P_{hetero}(I_t) P_{total}(I_{t-1}^*) \\ &= P_{total}(I_t) \end{aligned} \quad (9)$$

6 Experimental results

We present in this section some results obtained a soccer game captured by two synchronized stationary video cameras. The results illustrate the robustness of the tracking of moving objects across views and show its efficiency in tracking the multiple moving objects with partial and total occlusions.

In Figure 3.a, we present the performance of the proposed approach in a complex case. In the blue box of each frame, there are five players generating a large number of occlusions. The proposed method tracks separately every player and correct estimation of the bounding box is generated for each player. Figure 3.b and Figure 3.c illustrate the extracted trajectories. In Figure 3.b, 3D trajectories of the players are shown from a synthetic view. The corresponding 2D trajectories viewed from each camera are presented in Figure 3.c. It can be observable that the trajectory of each player is consistent and continuous all the time.

The registration of the multiple heterogeneous cameras and the object tracking across these cameras are illustrated in Figure 4. In Figure 4.a, the moving object is occluded by a large obstacle (a kiosk), but it is continuously tracked using the motion estimation model and the information integrated from the static view: provided by the re-projected position of two lower corners of the bounding box. Figure 4.b, illustrates the capabilities of the proposed KF motion model in tracking separately the two overlapping moving objects. It allows us to split large moving blobs into sub-components or independent moving objects using their appearance and velocity properties. Finally, the ability of continuously tracking while the moving camera is panning the scene is presented in Figure 4.c and 4.d.

7 Conclusion

We have presented a novel approach for continuous tracking of multiple objects with a large amount of occlusions across multiple stationary or heterogeneous cameras. As proposed in Section 2, detection of moving camera sequence is simplified using affine transformation, so that existing background-based modeling detection methods can be applied. All available views are pre-registered using the homography computed from the ground plane model. This allows to simultaneously tracking 2D and 3D objects' positions in the field. For the heterogeneous set of cameras, a spatio-temporal homography is used to register cameras. The modeling of blobs appearance along with its 2D and 3D, or a combinatorial motion provides a robust method that tracks separately every moving object albeit the large amount of occlusions.

Several issues have yet to be addressed such as expanding the joint probability model to other criteria such as spatio-temporal invariant shape descriptors, and adapting higher order estimation methods for introducing non-linearity into the measurement step that will improve the accuracy of the tracking.

Acknowledgements

This research was partially funded by the Advanced Research and Development Activity of the U.S. Government under contract MDA-908-00-C-0036.

References

- [1] A. Elgammal and L. S. Davis, "Probabilistic Framework for Segmenting People Under Occlusion", *In Proc. ICCV*, 2001.
- [2] A. Mittal and L. S. Davis, "M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene Using Region-Based Stereo", *In Proc. ECCV*, 2002.
- [3] G. Stein, "Tracking from Multiple View Points: Self-calibration of Space and Time", *IEEE Proc. CVPR*, pp. 521-527, 1999.
- [4] H. Zhang and J. Malik, "Learning a discriminative classifier using shape context distance", *IEEE Proc. CVPR*, Vol. 1, pp. 242-247, 2003.
- [5] J. Kang, I. Cohen and G. Medioni, "Continuous Multi-Views Tracking using Tensor Voting", *In Proc. of IEEE WMCV*, 2002.
- [6] J. Kang, I. Cohen and G. Medioni, "Continuous Tracking Within and Across Camera Streams", *In Proc. of IEEE CVPR*, 2003.
- [7] I. Cohen and H. Li, "Inference of Human Postures by Classification of 3D Human Body Shape", *In Proc. of IEEE IWAMFG*, 2003.
- [8] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale and S. Shafer, "Multi-camera Multi-person Tracking for EasyLiving", *In IEEE proc. of VS 2000*, 2000.
- [9] Q. Cai and J.K. Aggarwal, "Automatic Tracking of Human Motion in Indoor Scenes Across Multiple Synchronized video Streams", *IEEE Proc. ICCV*, 1998.
- [10] R. T. Collins, A. J. Lipton, Ho Fujiyoshi and T. Kanade, "Algorithms for Cooperative Multisensor Surveillance", *Proc. of the IEEE*, Vol. 89(10), pp. 1456-1477, Oct, 2001.
- [11] Y. Chen, Y. Rui, and T. S. Huang, "JPDAF Based HMM for Real-Time Contour Tracking", *IEEE Proc. CVPR*, Vol 1. pp. 543-550, 2001.



Figure 3. Experiment results on the soccer sequence. (a) Tracking and segmenting multiple players rushing towards the ball, (b) 3D trajectories of moving objects in a synthetic view, (c) 2D trajectories of moving objects viewed in each camera.

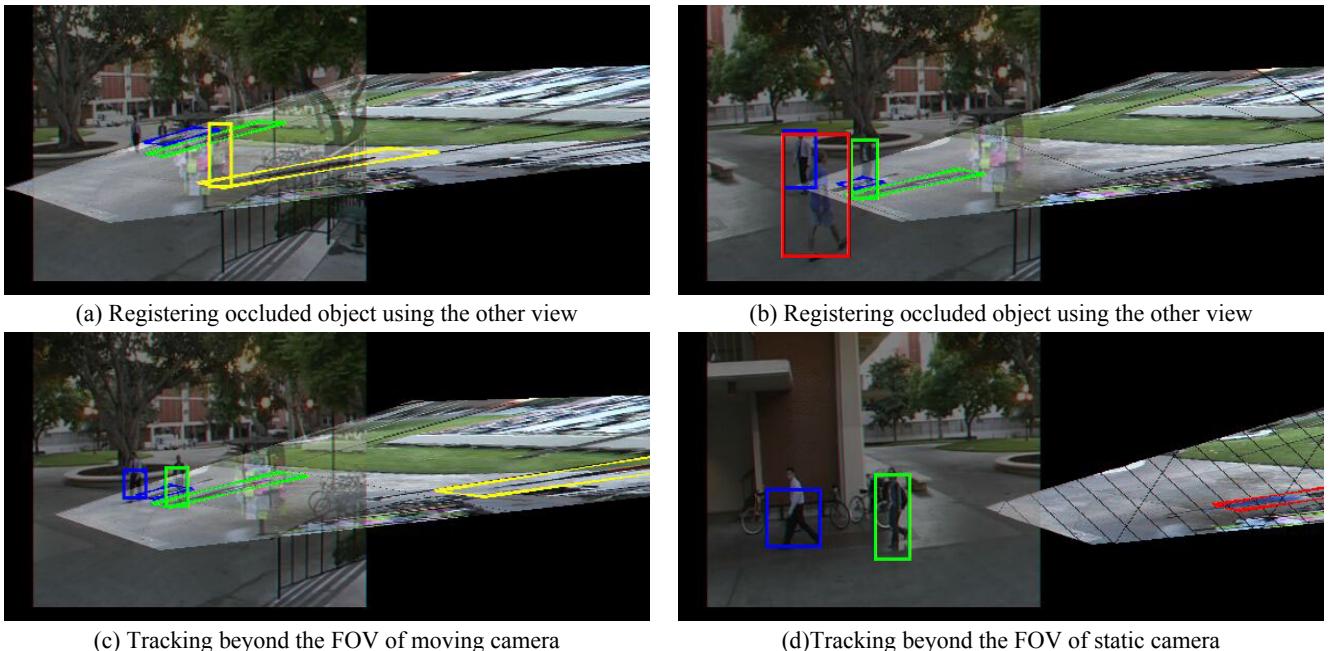


Figure 4. Detecting and tracking multiple objects across stationary and moving cameras

