A Two-stage Multi-view Analysis Framework for Human Activity and Interactions

Sangho Park
Computer Vision and Robotics Research Lab.
University of California, San Diego
La Jolla, CA 92037
parks@ucsd.edu

Mohan M. Trivedi
Computer Vision and Robotics Research Lab.
University of California, San Diego
La Jolla, CA 92037
mtrivedi@ucsd.edu

Abstract

This paper presents a new framework for a multi-stage multi-view approach for human interactions and activity analysis. The analysis is performed in a distributed vision system that synergistically integrate track- and body-level representations across multiple cameras. Our system aims at versatile and easily-deployable system that does not require careful camera calibration. Main contributions of the paper are: (1) context-dependent camera handover for occlusion handling, (2) switching the multi-stage analysis between track- and body-level representations, and (3) a hypothesis-verification paradigm for top-down feedback exploiting spatio-temporal constraints inherent in human interaction. Experimental evaluation shows the efficacy of the proposed system for analyzing multi-person interactions. Current implementation uses two views, but extension to more views is straightforward.

1. Introduction and Motivation

Analysis of multi-person interactions involving objects is an important research problem in computer vision for a wide range of potential applications: video surveillance, security enforcement, event annotation, motion analysis in sports, etc. Multi-person interaction raises particularly difficult issues in computer vision: occlusion between objects and body deformation during interaction.

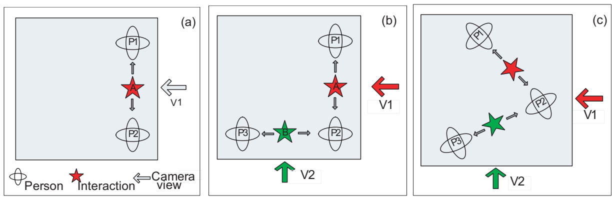

Figure 1. Top-down view diagrams for multi-view analysis of human interactions.

Fig. 1 illustrates multi-person interaction situations where the two-stage multi-view analysis would benefit. A single-camera system (Fig. 1 (a)) with viewing direction $V1$ may be sufficient for monitoring the two-person interaction A between persons $P1$ and $P2$, given the imaging condition is appropriate (i.e., with the viewing direction $V1$ orthogonal to the *interaction plane* that spans $P1$, A, and $P2$.) If the interaction plane is not perpendicular to the viewing direction, however, the single-camera based monitoring gets more difficult due to the occlusion and the change of appearance. With more than two persons involved (Fig. 1 (b)), a multi-view system may be inevitable even in the optimal viewing conditions; i.e., the viewing-directions, $V1$ and $V2$, are optimal for monitoring the interactions, A and B, between the persons $P1$ and $P2$, and $P2$ and $P3$, respectively. As the persons move around (Fig. 1 (c)), the dynamic selection and coordination of multiple views gets important, which is a challenging problem in computer vision; The incorporation of multiple cameras requires data fusion from each camera. Main difficulties in the data fusion from multiple cameras includes the question of how to decide when and which camera inputs to fuse for 2D and 3D image analysis. involved,

An integrated understanding of human activity involving body deformation would require multiple levels of analysis; we consider two stages of detail: track-level and body-level analyses. At the track level, human activity is analyzed in terms of the tracks of moving Gaussian ellipses that encompass individual persons. At the body level, human activity is analyzed in more detail in terms of the coordinated posture and gesture patterns of the body parts such as upper body and lower body. Major challenges in the two-stage analysis includes the maintenance of coherence between the two analysis stages; How can a vision system switch differ-

IEEE Workshop on Motion and Video Computing (WMVC'07)
0-7695-2793-0/07 $20.00 © 2007 IEEE

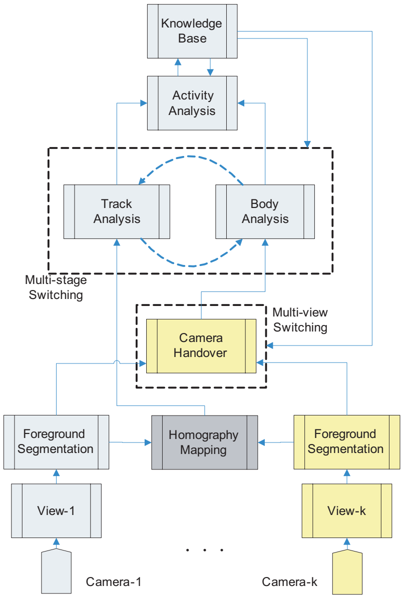

Figure 2. The overall system architecture.

ent analysis levels depending on the imaging quality under occlusion? The above two questions (i.e., multi-view fusion and two-stage fusion) may not be achieved by a simple uni-directional bottom-up or top-down vision process. Indeed, bidirectional process with some feedback mechanism is desirable, which would involve incorporation of top-down hypotheses about human interactions and bottom-up vision processes.

Majority of previous studies on human activity analysis have focused on track-level, single-perspective analysis. Reviews of general research on human motion understanding can be found in [1]. Majority of the approaches to behavior analysis are based on either body features from a single-view modality or composite features from multiple views such as histogram of 3D voxel [4], with calibrated cameras [8]or uncalibrated cameras [5]. A review of distributed surveillance systems can be found in [13]. Most of gesture recognition studies have aimed at learning isolated gestures of a single person with certain assumptions about camera configuration. Multi-view tracking and camera handover studies have not been actively related in activity recognition studies.

In this paper, we propose a new framework for the analysis of multi-person activity in a distributed vision system by a synergistic integration of the track- and body-level representations across multiple views. Main contributions of the paper are: (1) context-dependent camera handover for occlusion handling, (2) switching the multi-level analysis between track- and body-level representations, and (3) integration of data-driven bottom-up process and knowledge-driven top-down process for human activity understanding.

2. System Overview

Fig. 2 shows the overall system architecture. Light gray modules compose the basic single-view system, while the bright (yellow) modules compose the multi-view functionality. Dark gray module can work either in single- or multi-view modes, but more cameras can increase the overall accuracy. Currently, two cameras are used for synchronized views, which are foreground-segmented and combined to form a planar-homography map for 3D footage locations of the persons. The homography map is used for the track-level analysis. The camera handover searches for unoccluded person views for the body-level analysis. Both the track- and body-level analysis can be used for the activity analysis depending on analysis task. Semantic domain knowledge is incorporated with activity analysis, and provides constraints to other processing modules as feedback.

3. Multi-Level Analysis of Human Activity

The current system's analysis of human activity starts from foreground segmentation. We adopt a modified version of the codebook-based background model [6] to segment foreground regions. The background subtraction is followed by an 'attribute relational graph' based multitarget-multiassociation tracking (ARG-MMT) to segment and track multiple body parts simultaneously [11]. The multi-body tracking in ARG-MMT uses bounding boxes and 2D Gaussian ellipses to track the foreground bodies. As the people translate, the Gaussian parameters are updated along the sequence in a frame-by-frame manner. Updating these Gaussian parameters amounts to tracking the whole body translation of each person on the 2D image domain [12].

Our framework represents body motion at multiple layers: bounding box, 2D ellipse, and segmented body parts. Fig. 3 shows some example frames of the single-view ARG-MMT tracking results in which the segmented and tracked body parts are marked by distinctive artificial color at the body-parts layer which is embedded in the bounding-box layer in various imaging situations (The intermediate 2D-ellipse layer is omitted for clarity.) Occlusion during human interaction degrades the body part segmentation, while the coarse bounding box and ellipse are well maintained. Proper handling and estimation of body parts during occlusion from a single perspective is still an open question in computer vision. This motivates us to develop a synergistic two-level analysis framework and an adaptive mechanism to switch the track-level and the body-level analysis.

0-7695-2793-0/07 $20.00 © 2007 IEEE

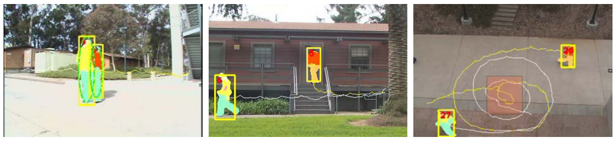

Figure 3. Tracking results at various sites with different camera configurations.

3.1. Track-Level Activity Modeling

We represent the individual track pattern Γ_i^k of the i-th person at time k in terms of the following features:

TRACK FEATURE Γ_i^k:

- $\Gamma_i^k = [P_i^k, d_{ij}^k, \dot{d}_{ij}^k, \|v_i\|, \angle v_i]$
- P_i^k : coordinates of the current track position in 2D space
- d_{ij}^k : relative distance between the i-th person and the most adjacent j-th person at frame k
- $\dot{d}_{ij}^k$: derivative of the relative distance
- $\|v_i\|$: track velocity magnitude
- $\angle v_i$: track velocity orientation

The above features are extracted from a least mean square based polynomial regression [14] curve of the track points computed along a moving window of size ρ seconds (currently, ρ = 1 second.) The track points are perspective-compensated by planar homography to unwarp the imaging artifact. The adjacency in the formulation d_{ij}^k is a predicate that represents whether the distance is within a certain proximity from other persons. The main interest in the track-level analysis includes the estimation of a moving person's speed, perimeter sentry for cautious or secured areas, the estimation of proximity between persons, etc.

3.2. Body-Level Activity Modeling

The track-level analysis can not handle the detailed body-level human activity patterns performed by stationary people: e.g., shaking hands, dancing, pushing, kicking, etc.

We formulate the body-level person activity in terms of a stochastic estimation of poses and gestures using Hidden Markov Models. The pose estimation starts by extracting the occupancy map (OM) of the person by overlaying a 9 × 10 grid on the foreground silhouette and counting the normalized histogram of the foreground pixels within each cell of the grid ranging [0, 1]. The occupancy map is bisected into upper body and lower body, to form a 45 dimensional feature vector that represents the individual person's upper and lower body silhouettes, respectively. Vector quantization using K-means clustering reduces the dimensions of the feature vectors; The K codewords of the clusters are trained with training data that spans various types of single person activity. A human gesture is represented by a sequence of the codewords, and recognized by HMMs.

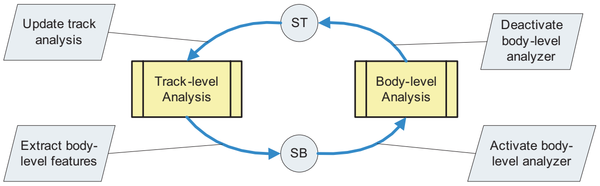
Update track analysis
ST
Deactivate body-level analyzer
Track-level Analysis
Body-level Analysis
Extract body-level features
SB
Activate body-level analyzer

Figure 4. Two-stage processes for activity analysis.

HMM-based approaches to activity recognition have been presented in [9]with different feature sets. We use independent sets of HMMs to represent the upper body gestures, lower body gestures, and torso translations, respectively. The assumption of independence between the individual HMMs dramatically reduces the size of the overall state space and the number of relevant joint probability distributions.

3.3. Switching Two Analysis Levels

The sensitivity of the body-level analysis is affected by many sources of uncertainty including occlusion, articulation, camera perspective, imaging noise, etc. In contrast, the track-level analysis is more robust across these conditions, and it is regarded as surveillance systems' baseline analysis which is always available.

The proposed algorithm switches to the body-level analysis whenever possible, and switches back to the track-level analysis whenever the body-part appearance quality degrades. This feedback-based iterative process is illustrated in Fig. 4: 'Switching to body level' (SB) occurs when reliable body information is available, and 'switching to track level' (ST) occurs when the body information gets unreliable. The body-appearance quality is evaluated by comparing the body-appearance fidelity feature F_j^k for person j at frame k with the learned previous frames. The individual features in Body-appearance fidelity F_j^k are obtained from the ARG-MMT process [11]. The pseudo-code for the context switching process is shown below (PSEUDO-CODE for switching algorithm.)

BODY-APPEARANCE FIDELITY FEATURE F_j^k:

- $F_j^k = [R_j^k, A_j^k, H_j^k, U_j^k, \theta_j^k]^T$
- R_j^k: Aspect ratio of the j-th bounding box B_j^k at frame k:

$$R_j^k = \frac{[Width\ of\ B_j^k]}{[Height\ of\ B_j^k]}$$

- A_j^k: Area of the j-th foreground pixels in B_j^k at frame k:

$A_j^k = \sum\sum_{(x,y)\in R} I(x,y)$

- H_j^k: Head ratio to height
- U_j^k: Upper-body ratio to height
- θ_j^k: Orientation of 2D Gaussian:

  $\theta_j^k = \frac{1}{2} \arctan\left(\frac{2\mu_{11}}{\mu_{20}-\mu_{02}}\right)$

  where $\mu_{pq} = \sum_x \sum_y (x - \bar{X})^p (y - \bar{Y})^q$ for $p, q = 0, 1, 2$.

```
PSEUDO-CODE for switching algorithm:
•Initialize system
 Scene modeling:
  – Perspective compensation by planar homography
  – Person footage detection
 Body modeling:
  – Gaussian mixture model to classify pixel color.
  – Blob generation by region growing wrt. color similarity.
  – Initialize / update the body-parts model:
    (i.e., head, upper- and lower- body)
•Do switching:
  – Compute new body-appearance fidelity feature $F_j^k$
  – If $F_j^k$ is over a threshold:
    Apply a body-level analysis using the HMMs
  – Otherwise:
    Apply track-level analysis using the ellipse-based tracking
```

4. Multi-View Analysis of Human Activity

The size of a pedestrian's appearance in wide-view surveillance video changes systematically according to the distance from the camera to the person. The aspect ratio of pedestrian appearance under the camera configuration with large inclination angle also changes systematically with the degree of camera inclination. In such situations, the track-level analysis will be always available, whereas the body-level analysis may be difficult due to degraded image appearance. A solution is to use multiple views for camera handover with different perspectives. The multi-view analysis is also useful to deal with occlusion that usually occurs in narrow field-of-view imaging conditions.

4.1. Region-based Homography Binding

Planar homography [3] can be used to locate the object's footage position on the world coordinate system. A homography matrix H maps corresponding points between different image coordinate systems. We use the 4-point algorithm [3] to compute the homography matrix H. The 4 points are selected from shared image corners or by having a person to walk around the shared region.

We map points in view 1, P_1, and points in view 2, P_2, to the corresponding points in the virtual top-down view, P_1^v and P_2^v, by homography matrices H_1^v and H_2^v, respectively, as follows:

$$P_1^v = H_1^v P_1, P_2^v = H_2^v P_2$$

Multiple views of the same object are projected by planar homography. Planar homography constraint assumes all the pixels lie on the same plane (i.e., the ground plane in 3D world.) Pixels that violate this assumption result in mapping to a skewed location on the projection plane. By intersecting multiple projection maps of the same object, we can estimate the object's common footage region that observes the assumption.

4.2. View Switching for Best-view Selection

Planar homography map provides view-independent coordinate values of the footage location of each person. However, it does not necessarily mean that the best-view selection is obtained, since the best view depends on camera perspective and individual person's body posture. The 'best' view also depends on recognition tasks; for example, side view would be better than front view for recognizing *pointing* interaction, while front view may be better than side view for face recognition. This finding leads us to a hypothesis-verification paradigm for best-view selection in a top-down manner.

The current system maintains two versions of best-view selection process: (1) estimation of separation (i.e., dispersedness) between persons and (2) task-dependent selection of persons of interest in specific torso orientation.

5. Human Interaction Modeling

A gap exists between geometric information obtained from images and semantic information contained in conceptual terms [7]. It is necessary to associate visual features with concepts and symbols to build event semantics of a person's activity. Our representation of multi-person activity is based on an event hierarchy introduced in [10] that is composed of instantaneous *poses*, body-part *gestures*, single-person *action*, and multi-person *interaction*.

Event representation may involve *interaction* between multiple sub-events. We represent the interaction of two events with respect to their spatial, temporal, and logical relations by using a predicate calculus. We adopt Allens interval temporal logic [2] to represent temporal relations between two events. Spatial relation of two events is represented by the spatial proximity of the events. Logical relations represent useful domain knowledge and constraints.

HMM-based event recognizer is used for training and testing of events in the give scene. Each object is associated with its own activity status represented by the operation triplet

0-7695-2793-0/07 $20.00 © 2007 IEEE

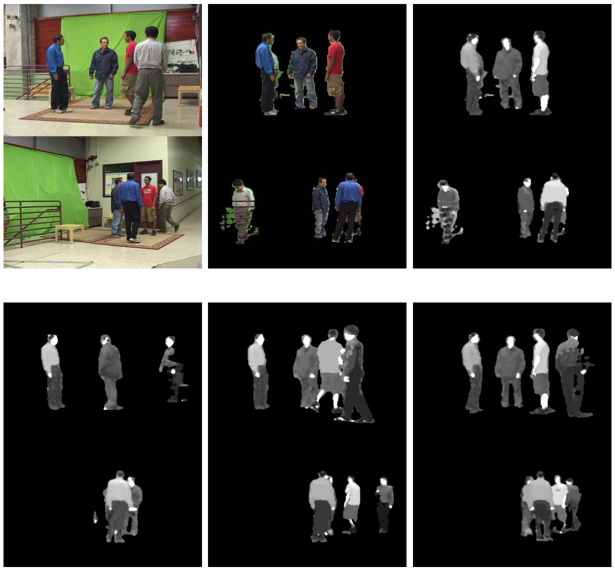

Figure 5. Multi-person activity situation viewed from two quasi-orthogonal perspectives.

6. Experimental Study

6.1. Experiment on Multi-view Switching

Fig. 5 shows example sequence of multi-view switching that selects the best view of separate persons. The 1^{st} image pair in the 1^{st} row shows two views from camera A (upper image) and B (lower image). The 2^{nd} and 3^{rd} image pairs show foreground segmented images and body part segmetnation results at frame 1828, respectively. The 2^{nd} row shows example frames of camera handover instances (i.e., cameras A → B → A) for best-view selection. (The frame numbers are 1398, 1587, and 1698, respectively, at 30 frames per second speed.)

6.2. Experiment on Two-stage Analysis

Fig. 6 show an example of two-stage analysis of human interaction: the track- and body- levels. The example sequence of *passing* interaction in Fig. 6 shows that person ID-5 in the middle goes upstairs to knock the door, and person ID-6 passes by, and person ID-5 goes downstairs to exit. The 2^{nd} row shows the corresponding raw frames. The 3^{rd} row shows the track-level analysis results in the image domain (i.e., XY plot) and the more detailed spatio-temporal domain (i.e., XYT plot), respectively. The track-level analysis shows each person's body translation; person ID-5 enters from right, goes upstairs, then exits to the right side, and person ID-6 traverses from the left to the right side.

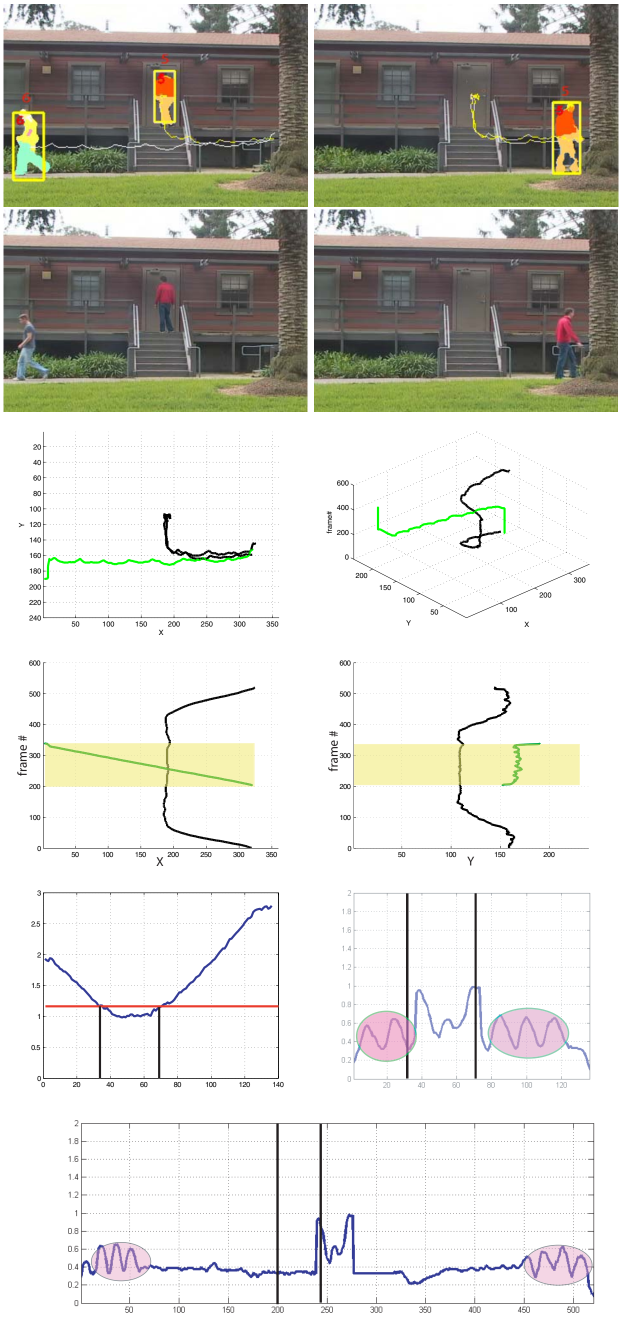

Figure 6. Two-stage analysis of *passing* interaction.

The 4^{th}-6^{th} rows show more detailed body-level analysis. The XT- / YT- plots (4^{th} row) show the overlapping period of the two persons' existence denoted by the gray (yellow) regions. The following analysis is based on this period. Euclidean distance D of the two persons' centroid is normalized by their average height (i.e., 'height-normalized'),

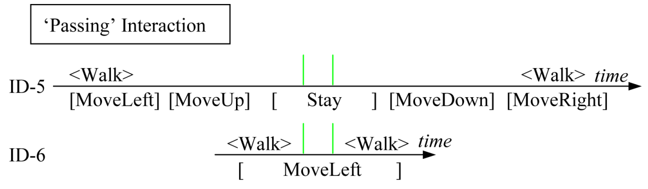

Figure 7. Semantic description of the *passing* interaction in Fig. 6.

and plotted along the timeline in abscissa (i.e., 'TD plot') on the 5^{th} row (1^{st} image.) The horizontal bar located at the interaction-distance threshold (i.e., the ordinate value of 1.2) in the TD plot provides time indices of the interaction; that is, the period of the interaction (1.2 seconds) is obtained by the two vertical projection lines onto the time dimension (i.e., frame number in abscissa.) The next two plots show the plots of height-normalized lower-body width of person ID-6 and ID-5, respectively. The overlayed ellipses denote the correctly classified patterns as 'walking' by the leg HMMs. The two-stage switching mechanism identified the erratic distortion of the body-appearance fidelity feature F (in Section 3.3) due to the occlusion, and nullified the HMM output as invalid.

The track-level and body-level analysis results are integrated into the semantic description with detected activities aligned along a common timeline. Fig. 7 represents the two-stage analysis for the *passing* interaction in Fig. 6: the track- and the body- level descriptions below and above the timeline, respectively. The period of proximal interaction is marked with the gray (green in color) vertical bars on the timeline. The semantic-level description in Fig. 7 provides users with intuitive and concise summary of events; multi-person *interaction* is obtained by focusing on the period of proximity between persons, while the description of individual *action* is available along the entire timeline.

7. Conclusions

We have presented a two-stage multi-view analysis framework for human activity and interactions. The analysis is performed in a distributed vision system that synergistically integrate track- and body-level representations from multiple views. Main contributions of the paper are: (1) context-dependent camera handover for occlusion handling, (2) switching the multi-level analysis between track- and body-level representations, and (3) integration of data-driven bottom-up process and knowledge-driven top-down process for human activity understanding. Experimental evaluation shows the efficacy of the proposed system for analyzing multi-person interactions. Current implementation uses 2 cameras, but extension to more cameras is straightforward. We are currently working on analysis of object-involved interactions among persons and on video database query method that exploits the advantage of multiple views.

Acknowledgments

This research was supported in part by US DoD Technical Support Working Group (TSWG) and by NSF RESCUE ITR-Project.

References

[1] J. Aggarwal and Q. Cai. Human motion analysis: a review. *Computer Vision and Image Understanding*, 73(3):295–304, 1999.

[2] J. F. Allen and G. Ferguson. Actions and events in interval temporal logic. *Journal of Logic and Computation*, 4(5):531–579, 1994.

[3] A. Criminisi, I. Reid, and A. Zisserman. A plane measuring device. *Image and Vision Computing*, 17(8):625–634, 1999.

[4] K. S. Huang and M. M. Trivedi. 3d shape context based gesture analysis integrated with tracking using omni video array. In *Proc. IEEE Workshop on Vision for Human-Computer Interaction (V4HCI)*, San Diego, USA, June 2005.

[5] S. M. Khan and M. Shah. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *ECCV*, 2006.

[6] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis. Real-time foreground-background segmentation using codebook model. *Real-Time Imaging*, 11, 2005.

[7] A. Kojima, T. Tamura, and K. Fukunaga. Textual description of human activities by tracking head and hand motions. In *Int'l Conference on Pattern Recognition*, volume 2, pages 1073–1077, 2002.

[8] A. Mittal and L. S. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. 51(3):189–203, 2003.

[9] N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. In *Proc. IEEE Int'l Conference on Multimodal Interfaces*, pages 3–8, 2002.

[10] S. Park and J. Aggarwal. Semantic-level understanding of human actions and interactions using event hierarchy. In *IEEE Workshop on Articulated and Nonrigid Motion*, Washington, DC, USA, 2004.

[11] S. Park and J. Aggarwal. Simultaneous tracking of multiple body parts of interacting persons. *Computer Vision and Image Understanding*, 102(1):1–22, 2006.

[12] S. Park and M. M. Trivedi. A track-based human movement analysis and privacy protection system adaptive to environmental contexts. In *IEEE Int'l Conference on Advanced Video and Signal based Surveillance*, Como, Italy, 2005.

[13] M. Valera and S. Velastin. Intelligent distributed surveillance systems: a review. *IEEE Proceedings Vision, Image and Signal Processing*, 152(2):192–204, 2005.

[14] E. Williams. *Regression Analysis*. Wiley, New York, 1959.