# Definition, Detection, and Evaluation of Meeting Events in Airport Surveillance Videos

Sung Chun Lee, Chang Huang, and Ram Nevatia

University of Southern California, Los Angeles, CA 90089, USA
sungchun@usc.edu, huangcha@usc.edu, nevatia@usc.edu

## Abstract

Detection of events in the surveillance video selected for TRECVID 2008 is an extremely difficult task. It was not possible for us to evaluate on a variety of events and the entire length of the dataset. We participated in an exploratory task and selected a single “people meeting” event. This event was selected due to its frequency, importance and difficulty. It is actually a collection of similar tasks as there can be several styles of meeting: for example two people coming towards each other, one person waiting for others, a person joining an existing group etc. Difficulty of the task can vary based on the density of the crowd in the scene at the moment and the extent to which the participants are occluded. It is not too useful to just combine the results of all these conditions and simply average; we selected video segments where participants are partly occluded and the crowd density is "medium". We first detect and track pedestrians using an existing tracking method; in fact, tracking in this complex environment is the strongest challenge for the task. Then, we detect meeting events based on analysis of the trajectories. We present results which include the precision and recall rates and some graphical results.

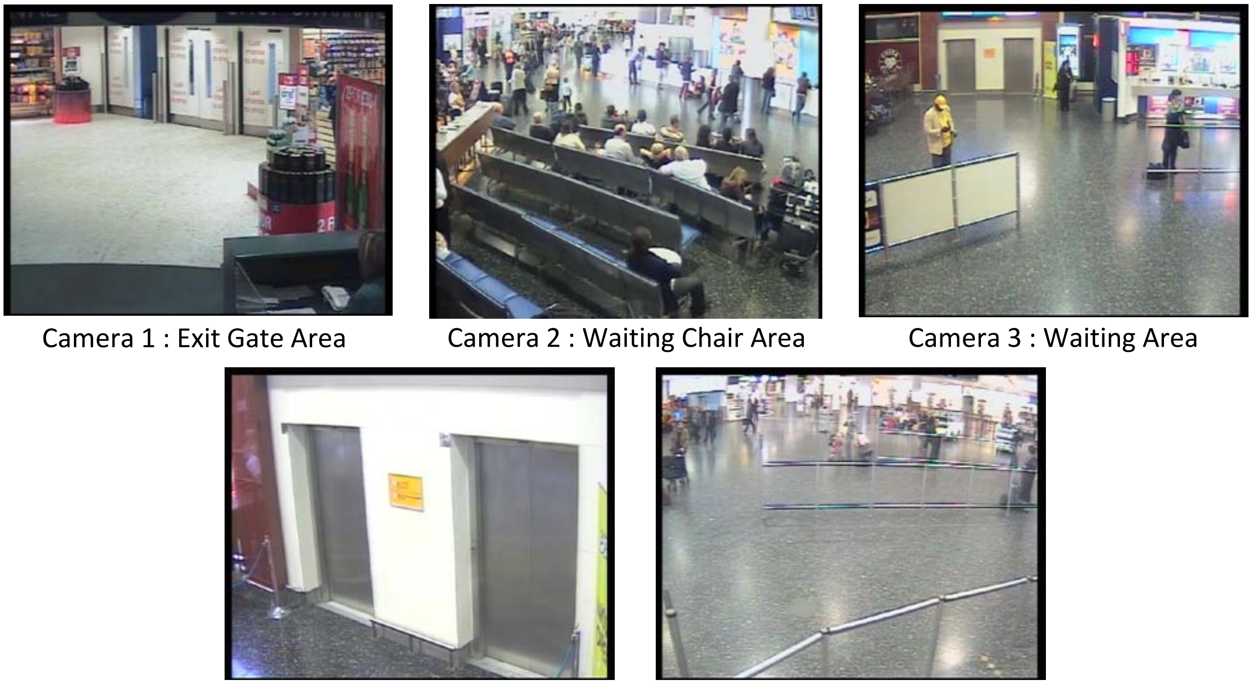
## 1. Introduction

TRECVID 2008 has selected the task of event detection in airport surveillance videos [1]. The source data comprise about 100 hours (10 days * 2 hours per day * 5 cameras) of video obtained from Gatwick Airport surveillance video data. A number of events for this task were defined. Our approach to event recognition may be described as a *bottom-up* approach where we first detect and track people and other objects and then infer the activities from the tracks. We believe that this approach is much more robust than the *top-down* approaches using global properties of a video; however, the processing is slow and results are highly dependent on the accuracy of the tracks. The airport data is judged, by us, to be beyond the state-of-art for bottom-up methods to analyze sufficiently well to constitute a meaningful evaluation. Also, the process is computation intensive and it was not practical for us to analyze 100 hours of video.

Given the above considerations, we chose to participate in a freestyle evaluation task that allowed us to select events of interest and relevant portions of the data that can provide insights into strengths and weaknesses of our approach. We selected the 'meeting event detection' task as such events are frequent in the airport video and their detection is of reasonably high complexity. We further selected data where the event exhibits as being of medium complexity.

## 2. Meeting Event Data Selection

In TRECVID 2008 airport surveillance video data, there are five cameras from different location in the airport as shown in Figure 1. We excluded three cameras (Camera 1, Camera 2, and Camera 4) that contain few meeting events. We formed a group of 8 people to view the 2008 TRECVID video data set of 'Camera 3' and 'Camera 5' and to collect video segments that contain meeting events.



Camera 1 : Exit Gate Area

Camera 2 : Waiting Chair Area

Camera 3 : Waiting Area

Camera 4 : Elevator Area

Camera 5 : Passage Area

Figure 1. Five camera video scenes of TRECVID 2008 data.

As a result, we collected 242 video segment clips of meeting events. We subjectively analyzed these video clips and categorized the hardness of the scene by the following attributes:

1) Occlusion of participants (None / Partial / Full)
2) Resolution (image size) of participants (Low / Medium / High)
3) Crowdedness of the scene (Low / Medium / High)
4) Number of people in meeting event (two / three / four / five / six)

Figure 2 shows the distribution of the collected video segments. Partial occlusion and medium resolution of two participants in middle level crowded scene is the most common event among the collected data.

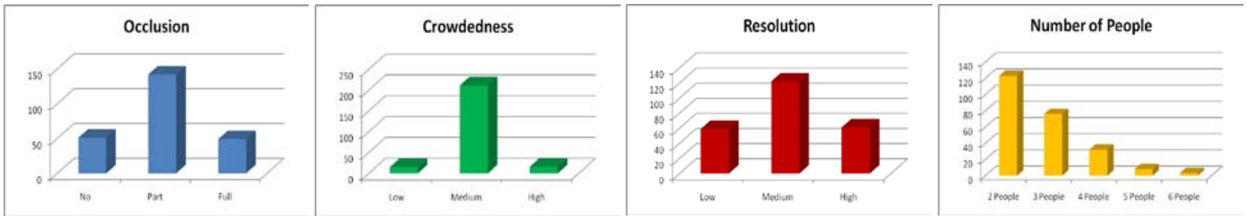


Figure 2. Meeting event frequency analysis per each category.

As next step, we integrated inter-category classes to estimate the complexity of the data.  For example, '*no occlusion / low crowdedness / high resolution / 2 people*' is the easiest case and '*full occlusion / high crowdedness / low resolution / 6 people*' is the most difficult case. Figure 3 shows the meeting event frequency graph according to its hardness.

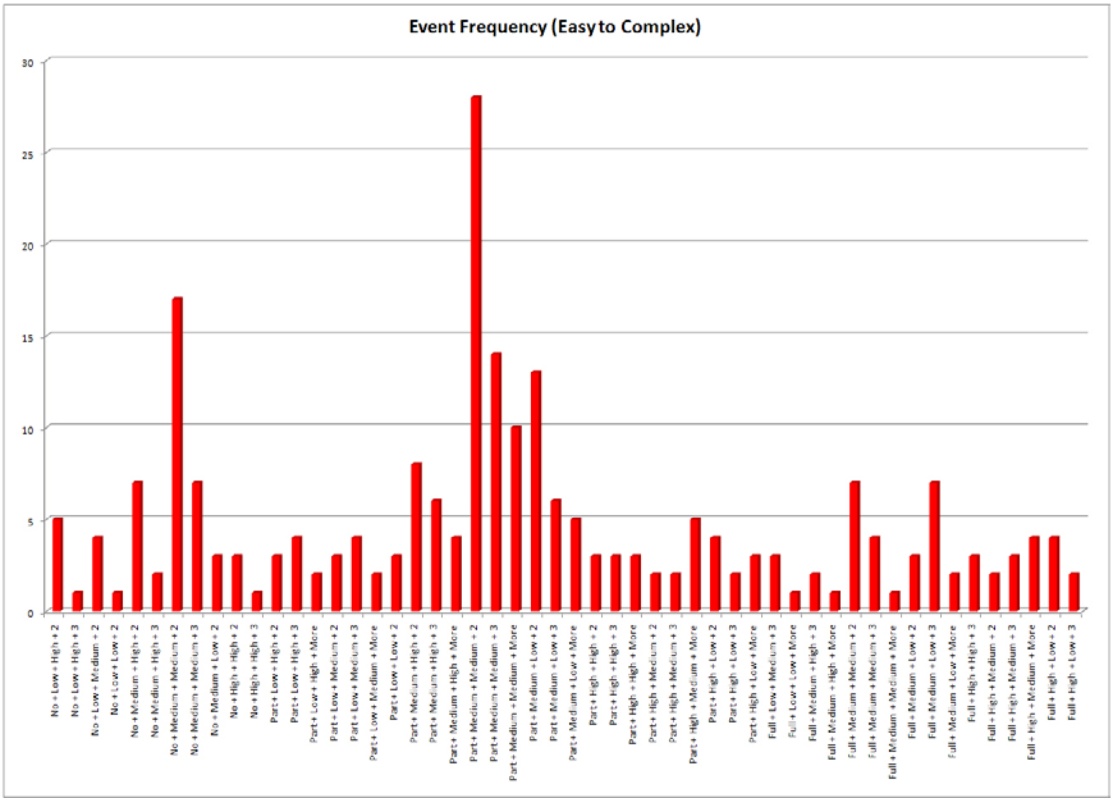


Figure 3. Meeting event frequency analysis of inter category.

We chose 17 video clips from the most frequent cases (*part occlusion / medium crowdedness / medium resolution / 2 or 3 people*) which are middle levels of hardness, to test our proposed event detection method. A key frame of some of the selected video segments is shown in Figure 4.

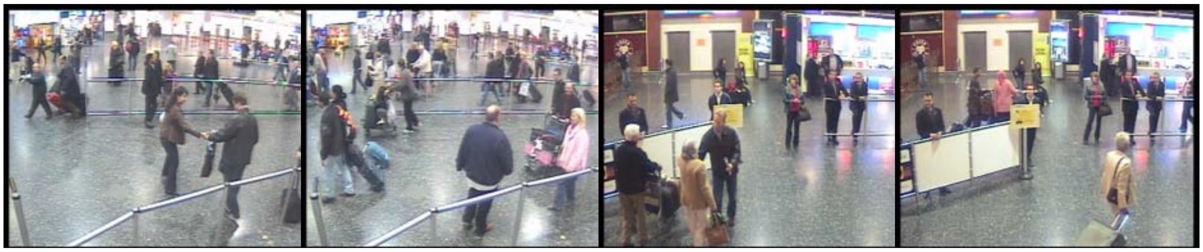


Figure 4. Examples of meeting events.

## 3. Tracking Humans

We detect and track people in the scene by using the method described in [2] (only a brief summary is given here). Pedestrian tracking is a process of locating upright standing or walking humans in videos, maintaining their identities through time, and finally obtaining their trajectories. In this paper, meeting events are detected mainly by analyzing pedestrian trajectories. Therefore, pedestrian tracking plays an important part in our meeting event detection system. However, it is really a difficult problem since meeting events usually happen in complex and crowded environments, where many pedestrians, moving in different directions are present simultaneously; many pedestrians have similar appearances and occlude one another and occlusions by other scene objects are also common. To obtain reliable pedestrian trajectories for meeting event detection task, we propose a detection-based hierarchical association method that is capable of robustly tracking multiple pedestrians under such challenging conditions.

Our approach generates pedestrian trajectories by means of progressively associating detection responses given by Wu-Nevatia pedestrian detector [3]. This hierarchical approach consists of three successive levels, of which each adopts different models and methods shown as Table 1.

| | **Motion Model** | **Appearance Model** | **Association Method** | **Scene Model** | **Coordinate** |
|---|---|---|---|---|---|
| Low Level | n/a | Raw | Direct Link | n/a | Image Plane |
| Middle Level | Dynamic | Refined | Hungarian | General | Image Plane |
| High Level | Dynamic | Refined | Hungarian | Specific | Ground Plane |

**Table 1. Models and methods used in different levels of the hierarchical associate framework.**

First, at the low level, short but reliable tracklets are generated by linking detection responses in neighboring frames. Second, at the middle level, the short tracklets obtained at the low level are progressively associated into longer and longer tracklets. Each round of association is formulated as a MAP problem that considers not only initialization, termination and transition of tracklets but also hypotheses of tracklets being false alarms. Finally, at the high level, a scene structure model, including three maps for entries, exits and scene occluders respectively, is introduced as a hidden state of the system. We estimate this scene structure model and compute the optimal tracklet associate in an EM manner. In the E-step, the scene model is estimated by means of Bayesian inference from the already obtained tracklets; afterward in the M-step, the long-range tracklet association is performed with the help of the scene knowledge based reasoning to reduce trajectory fragmentation and prevent possible identity switches. The optimized tracklets at the high level are output as the final trajectories for pedestrians in the video.

## 4. Meeting Event Detection

Given pedestrian tracklets, we scan the frames of a video segment four times: the first scan is for detecting meeting event candidates, the second one is for tracklet association, the third scan is for selecting the meeting events, and the fourth scan is to merge any related meeting events among pedestrians. The first and third scans are identical processes as we check only two pedestrian tracklets at a time to decide whether they meet or not. In the fourth scan, we merge the detected meeting events with two pedestrians, which have a common pedestrian in their meeting event. The meeting event detection method (the first and the third scans) will be described in Section 3.3.1 and the tracklet association method will be explained in Section 3.3.2 followed by merging process in Section 3.3.3.

## 4.1 Meeting Event Hypotheses Generation

Given track information, we hypothesize meeting events using a rule-based detection method. We define a meeting event when all of the following conditions are satisfied:

1) **Closeness Condition**
   Pedestrians should stay close to each other during the meeting. The distance between pedestrians should be less than $d_m$ , a closeness distance threshold; otherwise, they are judged to be just passing by.
2) **Encountering Condition**
   Pedestrians should be away from each other before they meet. A distance between pedestrians at appearing time should be larger than $d_m$ . Otherwise, they have already met each other and we are not observing a meeting event.

We calibrate the cameras of the test video data using the method of [4]; this enables us to use the ground plane information and estimate 3-D distances.

Figure 5 shows some meeting event detection results. More detailed discussion will be described in the next Evaluation session.

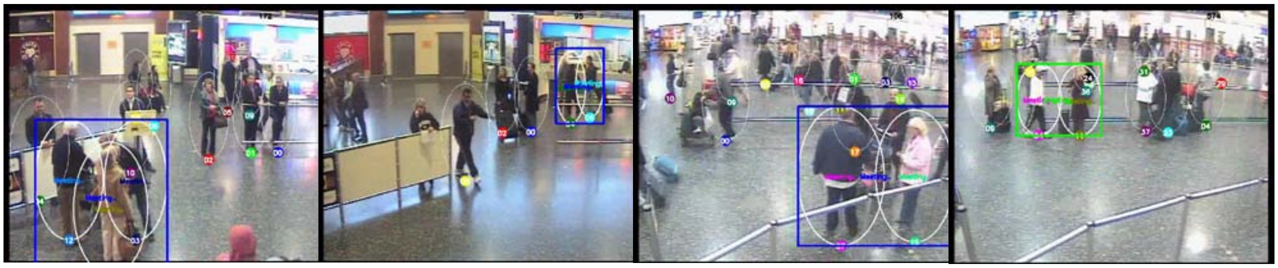


Figure 5. Results of meeting event detection (ellipses are tracklets and rectangles indicate the location of the detected meeting events).

## 4.2 Event-based Tracklet Association

There still exist many missing or broken pedestrian tracklets due to occlusion caused by the meeting event itself. We first search all possible candidates of meeting event throughout the video sequence. When there is a meeting event hypothesis and one of pedestrian tracklets has disappeared, it suggests an occlusion case. In this case, we estimate the speed and moving direction of the missing object and we predict the future position of the missing object. We match color histogram information of the missing object with any new objects around the predicted location using a mean-shift method [5]. If a good match is found, we connect the new tracklet with the previously terminated one. We interpolate the location of the person between the end of the missed tracklet and the newly starting tracklet and insert the interpolated positions in the occluded frames. Then, we detect actual meeting events which are described next.

### 4.3 Merging of Related Meeting Candidates

The meeting detection in the first and the third scans, is focusing on only two pedestrian tracklets. We find any detected meeting events that have common pedestrian tracklet and merge those two events.

# 5 Meeting Event Detection Evaluation Results and Discussion

We next describe our evaluation methodology and the results of our system.

## 5.1 Evaluation Method

We annotated the meeting events ourselves. The annotator draws a rectangle where a meeting event is taking place and keeps track of the meeting event throughout the frame sequences. Two annotators worked separately on creating meeting event ground truth data. Afterwards, they discussed their own decisions to come up with a common annotation to define the ground truth data. Our system draws the rectangles where two or more human tracks meet as shown in Figure 5.

Rather than comparing events frame-by-frame, we evaluate them event-by-event. If a ground truth rectangle overlaps with the system output in any frame, we define it to be a correct detection. If there is no overlapping system output within the ground truth tracking rectangle, we define it as a missed detection. If there is no ground truth but there is a system output, then it is a false alarm.

## 5.2 Evaluation Results and Analysis

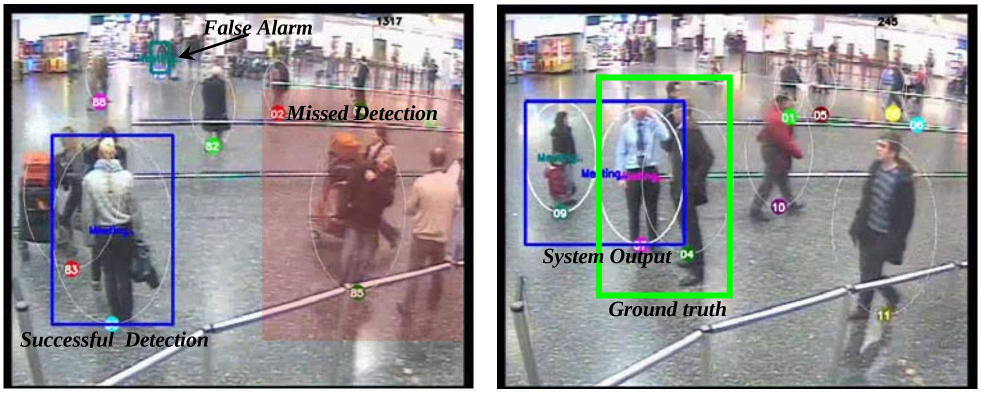
From 17 selected video segments, our annotators found 22 meeting events (ground truth) and our system found 20 events that match with ground truth, 2 missed detections and 29 false alarms. The recall rate and precision rate are shown in Table 2.

| **The Number of Ground Truth** | **The Number of Correct Detection** | **The Number of Missed Detection** | **The Number of False Alarm** | **Recall Rate**<br>$\left(\frac{\text{Number of Detection}}{\text{Number of Ground Truth}}\right)$ | **Precision Rate**<br>$\left(\frac{\text{Number of Detection}}{\text{Number of Detection} + \text{Number of False Alarm}}\right)$ |
|---|---|---|---|---|---|
| 22 | 20 | 2 | 29 | 0.91 | 0.41 |

**Table 2. Detection result analysis.**

The most common cause of false alarms comes from “passing by” cases at the end of video segments, where the system considers it to be a meeting event due to lack of other information. Another common reason of false alarms is due to low resolution (small image size) of pedestrians that are far away from the camera as shown in the top right of Figure 6 (a); this causes errors in detection and tracking stage. Figure 6 (a) shows an example of successful detection, false alarm, and missed detection.

Figure 6 (b) illustrates an issue of our evaluation method. The system does not find the correct meeting event (green box: ground truth). Instead, the system detects a meeting event (blue box: system output) between wrong people where the actual meeting event is supposed to happen as shown in Figure 6 (b). Because the evaluation method does not count which people meet each other but just measures the intersection between the system output and ground truth, this is considered as a successful detection. This suggests the need for more complex evaluation criteria which would also need to be supported by more complex annotations (such as encoding the identities of the participants in addition to their locations).



(a) (b)

Figure 6. Meeting event detection result analysis.

## 6 References

[1] Smeaton, A. F., Over, P., and Kraaij, W., Evaluation campaigns and TRECVid. *In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, *ACM Press*, New York, NY, 321-330. DOI= http://doi.acm.org/10.1145/1178677.1178722, 2006.

[2] Huang, C., Wu, B., and Nevatia, R.: Robust Object Tracking by Hierarchical Association of Detection Responses, *To be published in the 10th European Conference on Computer Vision*, 2008.

[3] Wu, B., Nevatia, R.: Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors, *International Journal of Computer Vision*, 2007.

[4] Fengjun Lv, Tao Zhao, Ram Nevatia, Self-calibration of a camera from video of a walking human, *International Conference on Pattern Recognition (ICPR)*, 1:562-567, 2002.

[5] D. Comaniciu, V. Ramesh, and P. Meer, Real-time tracking of non-rigid objects using mean shift, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1:511-518, 2001.