

# Speaker Tracking in Seminars by Human Body Detection

Bo Wu, Vivek Kumar Singh, Ram Nevatia, and Chi-Wei Chu

University of Southern California  
Institute for Robotics and Intelligent Systems  
Los Angeles, CA 90089-0273  
{bowu|viveksin|nevatia|chuc}@usc.edu

**Abstract.** This paper presents evaluation results of a method for tracking speakers in seminars from multiple cameras. First, 2D human tracking and detection is done for each view. Then, 2D locations are converted to 3D based on the calibration parameters. Finally, cues from multiple cameras are integrated in an incremental way to refine the trajectories. We have developed two multi-view integration methods, which are evaluated and compared on the CHIL speaker tracking test set.

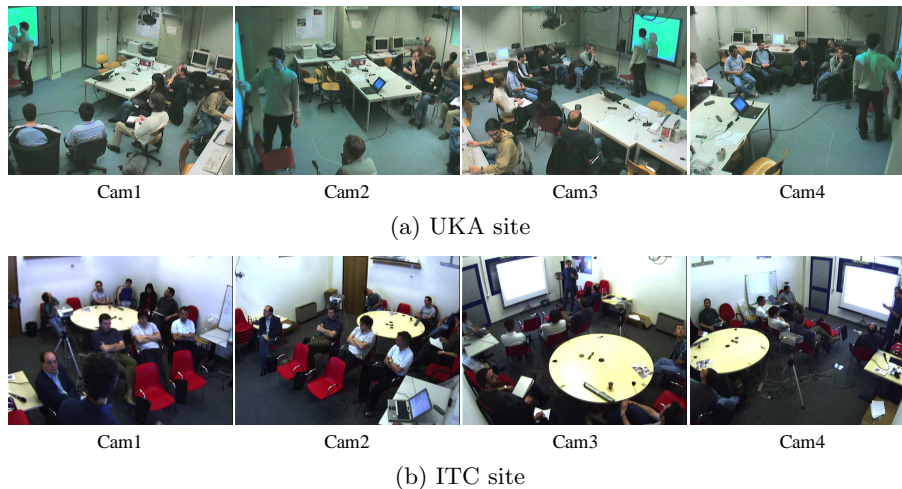
## 1 Task and Data Set

The task in this evaluation exercise is to track the 3D head locations of a speaker in seminars. In practice, only the ground plane projections of the 3D head locations are used to evaluate the performance. The test set contains 24 segments captured from the UKA site and 2 segments from the ITC site. For each segment, four side view cameras and one optional top-down camera are used to record the seminar. Camera calibration information including radial distortion, is provided. Each video contains about 4,500 frames. There are overall  $26 \times 4 \times 4500 = 468,000$  frames to process. The frame sizes of the UKA videos and the ITC videos are  $640 \times 480$  and  $800 \times 600$  respectively. The sampling rate of all videos is 30 FPS. Fig.1 shows some sample frames.

This task is made complex due to many reasons. First, the faces of the speakers are not always visible, so face or skin-color detection based methods can not be used in all cases. Second, the speaker does not move all the time and a clear scene shot is not available, hence moving object detection based on static/adaptive background modeling is difficult. Third, the scene is cluttered due to various scene objects, *e.g.* chairs and laptops.

Based on the observation that the speaker is usually the only person standing/walking during the seminar, we use a 2D multi-view human body detector [1] to locate the speaker frame by frame and track in 2D. Then the 2D trajectories are converted to 3D based on the camera calibration information. Finally the cues from multiple cameras are integrated to refine the trajectories.

The rest of the paper is organized as follows: Section 2 describes our tracking method; Section 3 shows the experimental results; and Section 4 sums up.



**Fig. 1.** Sample frames.

## 2 Methodology

We take the single frame human detection responses as the observation of human hypotheses. Tracking is done in 2D for individual views. The 3D head locations are calculated by triangulation or approximated with the 3D feet positions obtained from the calibration information and a ground plane assumption. Then cues from multiple cameras are integrated in an incremental way.

### 2.1 Multi-View Human Body Detection and Tracking

In the method of [1], four part detectors are learned for full-body, head-shoulder, torso, and legs. We only use the one for full-body here. Two detectors are learnt: one for the left profile view, and one for the frontal/rear view (the detector for right profile view is generated by flipping the left profile view horizontally). Nested cascade detectors are learned by boosting edgelet feature based weak classifiers, as in [1]. The training set contains 1,700 positive samples for frontal/rear views, 1,120 for left profile view, and 7,000 negative images. The positive samples are collected from the Internet and the MIT pedestrian set [2]. The negative images are general scene images collected from the Internet. The training sets are fully independent of the test sequences. For detection, the input image is scanned by all three detectors and the union of their responses is taken as the multi-view detection result.

The speaker is tracked in 2D by associating the frame detection responses. This 2D tracking method is a modified version of that in [3]. In [3], the detection responses come from four part detectors and a combined detector. To start a trajectory, an initialization confidence *InitConf* is calculated from  $T$  consecutive responses, which correspond to one human hypothesis, based on the cues from

color, shape, and position. If  $InitConf$  is larger than a threshold  $\theta_{init}$ , a trajectory is started. To track the human, first data association with the combined detection responses is attempted; if this fails, data association with the part detection responses is attempted; if this fails again, a color based meanshift tracker [4] is used to follow the person. The strategy of trajectory termination is similar to that of initialization. A termination confidence  $EndConf$  is calculated when an existing trajectory has been lost by the detector for  $T$  time steps. If  $EndConf$  is larger than a threshold  $\theta_{end}$ , the trajectory is terminated.

In this work, only the full-body detector is used. We do not use the combined detection [1] for partial occlusion reasoning explicitly, as the local feature based full-body detector can work with partial occlusion to some extent and occlusions are not strong in this data set. The tracker in [3] tracks multiple persons simultaneously; while the tracker in this work is designed to track a single person. Once a human trajectory is initialized, it prohibits the initialization of other trajectories. The result of the 2D tracker is a set of 2D trajectories which are temporally disjoint with each other. These trajectories share the same identity, *i.e.* they are considered corresponding to the same object. Fig.2 shows some sample frames of 2D tracking results.

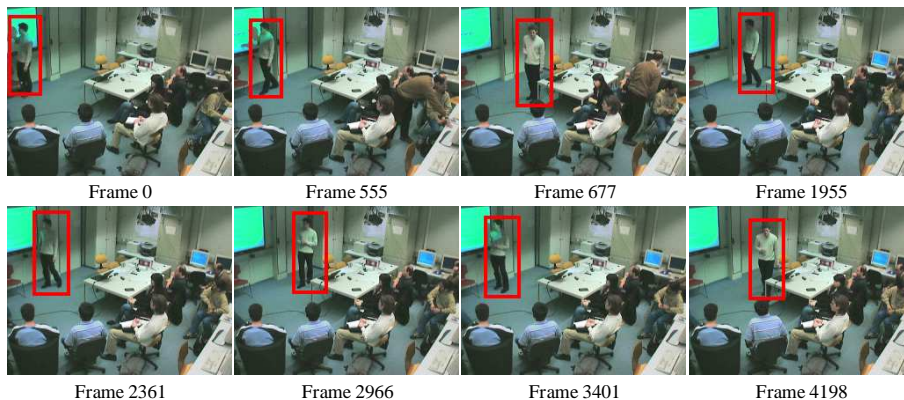


Fig. 2. 2D speaker tracking result.

## 2.2 Conversion from 2D to 3D

The 2D human detection and tracking only gives the rough 2D locations of the speaker. We need to extract the ground plane projections of the 3D positions of the speaker's head. We propose two methods to do this.

**Approximation by Feet Position** As we have the camera calibration information, the 3D feet positions can be calculated from the 2D pixel locations for

individual views based on an assumption that the speaker stands or walks on a ground plane. 3D feet positions are good approximation of the ground plane projections of 3D head positions. In practice, based on the human model of the positive training samples [1], we calculate the 2D feet positions from the rectangle-shaped detection responses, then project them to 3D space. Fig.3 illustrates the computation of the 3D feet positions.

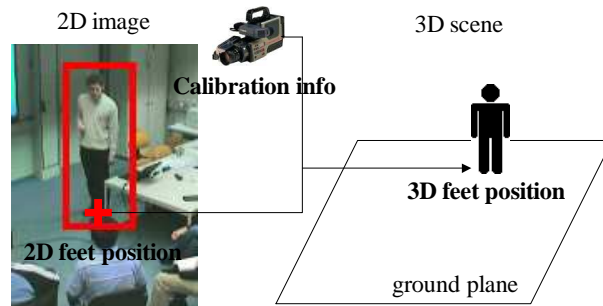


Fig. 3. Computation of 3D feet positions.

**Head Position by Triangulation** Similar to the case of 2D feet positions, based on the human model we can get the 2D head positions from the detection responses. Then we use a motion segmentation based method to further refine the 2D head positions. When the speaker is detected as moving, we search for the peaks of the foreground blobs within the response rectangles and take the peaks as the image positions of the head top. When the speaker is detected as being stationary, we just use the head positions calculated based on the human model. As the height of the speaker is unknown, we do triangulation from two views to get the 3D head positions. Fig.4 illustrates the computation of 3D head positions.

### 2.3 Integration of Multiple Cameras

For one segment, 3D trajectories are obtained from each camera. Partial occlusion of the speaker by the background or other persons may result in tracking errors. Also the speaker is not always visible from a single camera. In order to refined the 3D trajectories, we combine the tracking results from the individual cameras to form a multi-camera result.

Due to the errors in 2D tracking, the 3D trajectory may have some unnatural, sudden motions that we call *peaks*. We detect these peaks by thresholding the velocity of the trajectory. Denote by  $v_i$  the maximum magnitudes of the velocity of the  $i$ -th point,  $P_i$ , in the trajectory, and denote by  $d_i$  the overall translation of a sub-window  $W_i$  around  $P_i$ , *i.e.* the distance between the start point and

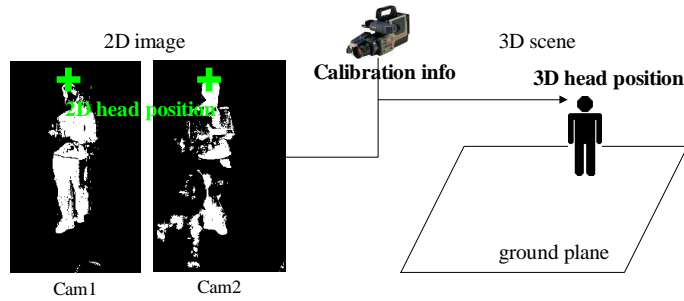


Fig. 4. Computation of 3D head positions.

the end point of  $W_i$ . If  $v_i$  is larger than a threshold  $\theta_v$  and  $d_i$  is smaller than a threshold  $\theta_d$ ,  $P_i$  is classified as peak and all points in  $W_i$  are removed from the trajectory.

This peak removal process reduces the false alarms in the tracking results but also creates some *gaps* (missed detections). Gaps may also be present if there is no detection from a single camera. We fill in these gaps by combining the trajectory information from all the cameras in an incremental way. We assign priorities to the individual camera outputs based on their accuracy on a small fraction of the development data. Starting from the highest priority camera, we remove peaks in the output 3D trajectory, then fill in these gaps by using the information from the next highest priority camera and so on. For the triangulation based method the initial 3D trajectory is generated from the best two cameras. This process is continued until all cameras have been used. Fig.5 illustrates the multi-camera integration.

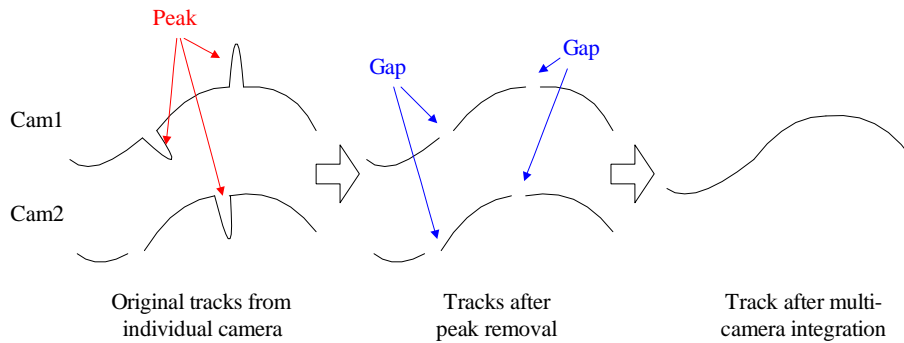


Fig. 5. Multi-camera integration.

### 3 Experimental Results

The formal evaluation process defines four metrics for the speaker tracking task [5]:

1. “Miss” represents the of missing detection rate;
2. “FalsePos” represents the false alarm rate;
3. Multiple Object Tracking Precision (MOTP) reflects the 3D location precision of the tracking level; and
4. Multiple Object Tracking Accuracy (MOTA) is the tracking accuracy calculated from the number of false alarms and the number of missed detections.

The first two metrics are for detection level and the last two for tracking level. If the distance between the tracked response and the ground truth is smaller than a threshold  $\theta_{pos}$ , it is considered to be a successful match; otherwise a false alarm. The default value of  $\theta_{pos}$  is 500mm. Table 1 lists the scores obtained with the default threshold, and Fig.6 shows the curves of MOTP and MOTA with different thresholds. The triangulation based method dominates the feet tracking based method, as the former locates the head directly. However the main advantage of the triangulation based method is the position accuracy; it can not improve the tracking level performance much when the threshold, *i.e.* the acceptable error, is close to one meter. Fig.7 shows the distribution of the tracking errors in 3D. Most of the errors are less than one meter, which is small compared to the size of the room.

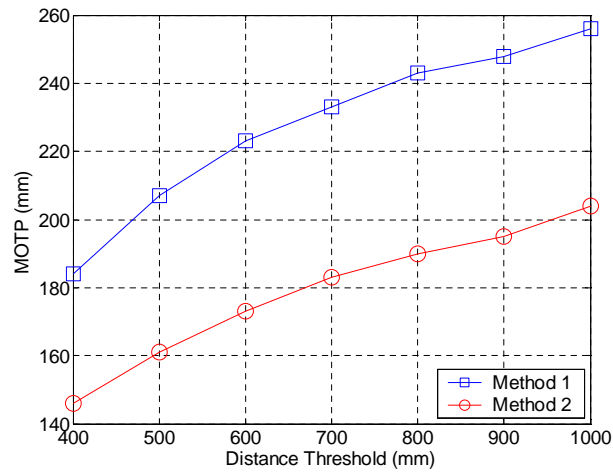
	Miss	FalsePos	MOTP	MOTA
Approximation by feet position	12.28%	12.22%	207mm	75.50%
Head position by triangulation	9.71%	9.65%	161mm	80.64%

**Table 1.** Evaluation scores with a default threshold of 500mm.

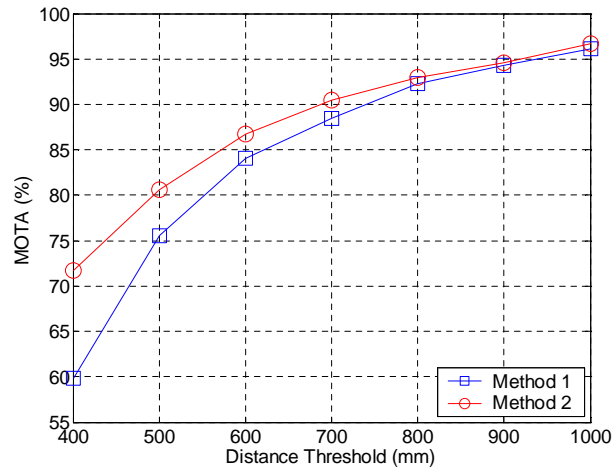
The speed of the system is about 2 FPS on a 2.8GHz Pentium CPU; the program is coded in C++ using OpenCV library functions; no attempt at code optimization has been made.

### 4 Conclusion and Discussion

We applied a fully automatic single human tracking method to the task of speaker tracking. The system achieves good performance on the test sequences. The comparative results between two multi-view integration methods shows that the triangulation based method has better accuracy.



(a) MOTP

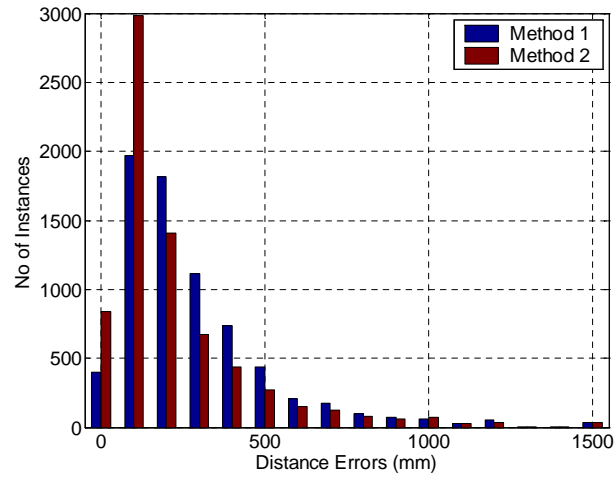


(b) MOTA

**Fig. 6.** Scores with different distance thresholds. (Method 1: approximation by feet positions; Method 2: head positions by triangulation)

Our current system does multi-view integration after the 2D trajectories are obtained. An alternative way is to do the integration after single frame detection, and then do tracking in 3D. This will remove some ambiguities at the detection level and make the tracking easier. We will explore this method in our future work.

**Acknowledgements:** This research was partially funded by the Advanced Research and Development Activity of the U.S. Government under contract MDA-904-03-C-1786.



**Fig. 7.** Error distributions. (Method 1: approximation by feet positions; Method 2: head positions by triangulation)

## References

1. B. Wu, and R. Nevatia. Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors. ICCV'05. Vol I: 90-97
2. C. Papageorgiou, T. Evgeniou, and T. Poggio. A Trainable Pedestrian Detection System. In: Proc. of Intelligent Vehicles, 1998. pp. 241-246
3. B. Wu, and R. Nevatia. Tracking of Multiple, Partially Occluded Humans based on Static Body Part Detection. To appear in CVPR'06.
4. D. Comaniciu, V. Ramesh, and P. Meer. The Variable Bandwidth Mean Shift and Data-Driven Scale Selection. ICCV'01. Vol I: 438-445
5. <http://www.clear-evaluation.org/>