

Evaluation of USC Human Tracking System for Surveillance Videos

Bo Wu, Xuefeng Song, Vivek Kumar Singh, and Ram Nevatia

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

Abstract. The evaluation results of a system for tracking humans in surveillance videos are presented. Moving blobs are detected based on adaptive background modeling. A shape based multi-view human detection system is used to find humans in moving regions. The detected responses are associated to infer the human trajectories. The shaped based human detection and tracking is further enhanced by a blob tracker to boost the performance on persons at a long distance from the camera. Finally the 2D trajectories are projected onto the 3D ground plane and their 3D speeds are used to verified the hypotheses. Results are given on the video test set of the VACE surveillance human tracking evaluation task.

1 Task and Data Set

The task in this evaluation exercise is to track the 2D locations and regions of multiple humans in surveillance videos. The videos are captured with a single static camera mounted a few meters above the ground looking down towards a street. The test set for the evaluation contains 50 sequences, overall 121,404 frames, captured from two different sites at various times. The frame size is 720×480 ; the sampling rate is 30 FPS. Fig.1 shows one shot of each site.



Fig. 1. Sample frames.

This task is made complex due to many reasons. The image appearance of pedestrians changes not only with the changing viewpoints and the clothing,

moving objects include not only humans but also vehicles, at the given resolution, detection of face pattern is infeasible and the scene is cluttered by many scene objects, *e.g.* trees and traffic signs.

We describe our method to overcome these difficulties in Section 2; Section 3 shows the experimental results; and Section 4 provides conclusion.

2 Methodology

We first detect person hypotheses in each frame, then we track them in 2D with a data association method. Shape based tracking is combined with a blob tracker to improve the performance. Based on computed camera parameters, the 2D trajectories are projected onto the 3D ground plane. The 3D speeds of the tracked objects are calculated and used to verify the hypotheses.

2.1 Shape based Human Detection and Tracking

We learn full-body detectors for walking or standing humans by the method proposed in [2]. Nested structure detectors are learned by boosting edgelet feature based weak classifiers. To cover different viewpoints, two detectors are learned: 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). The training set contains 1,700 positive samples for frontal/rear views, 1,120 for left profile view, and 500 negative images. The negative images are all of street scenes. The samples in training set are independent of the test sequences.

We do not use the combined detection in [2] for partial occlusion reasoning explicitly, as the local feature based full-body detector can work with partial occlusion to some extent and inter-human occlusions are not strong in this data set.

We constrain the search of humans around moving blobs. Motion is detected by comparing pixel color to an adaptively learned background model. If the proportion of the moving pixels within an image sub-window is larger than a threshold, θ_m , it is considered a candidate for human hypotheses and sent to the detector for further process; otherwise it is discarded directly. This reduces the search space of the human detector and prevents false alarms on static scene objects; however, it also prevents detection of static persons in the scene (in a real surveillance scenario, persons will always be expected to enter the scene at some time). Fig.2 shows some detection results.

Humans are tracked by forming associations between the frame detection responses. This 2D tracking method is a simplified version of that in [3], as only the full-body detector is applied. The affinity between a hypothesis and a responses is calculated based on cues from distance, size, and color. A greedy algorithm is used to associate the hypotheses and the detection responses. The automatic initialization and termination of trajectories are based on the confidences calculated from associated detection responses. To track the human, first data association with the full-body detection responses is attempted; if this fails,



Fig. 2. Sample detection results.

a color based meanshift tracker [4] is used to follow the person. Fig.3 show an example of shape based tracking.



Fig. 3. Examples of shape based human tracking results.

2.2 Motion based Human Detection and Tracking

The shape based detector does not work well on images of resolution where a person is less than 24 pixel wide, as in the case when the humans are far from the camera. Fig.4 shows some examples of missed detections.

We augment the shape based method with motion based blob tracking. Taking the motion detection results as input, we apply some morphological operations to connect foreground pixels to generate motion blobs. For simplicity, we



Fig. 4. Examples of missed detections. (The persons marked by the red arrows are detected as moving blobs but not found by the human detectors.)

model moving objects as rectangles. Each object is associated with an appearance model and a dynamic model. At each new frame, we predict the object's position with its dynamic model. Appearance model is used to distinguish objects when they are merged. A new object is created when a blob has no match with current hypotheses. A track ends when it has no blob match for more than a set number of frames. However, if multiple objects are merged in one blob from the beginning to the end, the blob tracker can not segment them. The moving objects with relatively small size are classified as pedestrians; the others to be vehicles. Fig.5 shows an example of motion based tracking.



Fig. 5. Examples of motion based human tracking results.

2.3 Combination of Shape and Motion based Approaches

We use an integration method to combine the shape based tracking and the motion based tracking. For each human track segment h_s from shape based tracking, we search for the motion blob track segments h_m which have large overlap with

h_s . We then merge the motion blob track segments h_m with the human track segment h_s . This combination increases the accuracy of the trajectories. Fig.6 shows an example of the combination.



Fig. 6. Combination of shape based and motion based tracking.

2.4 Verification by 3D Speed

False alarms of the detection system usually appear in cluttered or highly textured areas. Some of these false alarms are persistent ones, from which false trajectories may be generated. See Fig.7 for some examples.



Fig. 7. Examples of false alarm trajectories.

We use speed to discriminate humans from vehicles (vehicles can move slowly but human speed is limited). The image speed, however, depends on the position of the object in the image (faster motion near the camera). One approach could be to learn such speed patterns. We instead infer camera calibration parameters from observed motion by using approach proposed in [1] (this requires interactive processing in the tracking stage). Based on the calibration parameters and the assumption that objects are moving on a known ground plane, we project the 2D image locations onto the 3D ground plane, and calculate the 3D speeds of all tracked objects. If the average speed of a hypothesis is lower than a threshold, θ_{speed} , the hypothesis is accepted as human; otherwise, it is rejected.

3 Experiments

We ran our system on the 50 test sequences. The formal evaluation process defines four main metrics for the human tracking task [5]:

1. Multiple Object Detection Precision (MODP) reflects the 2D location precision of detection level;
2. Multiple Object Detection Accuracy (MODA) is the detection accuracy calculated from the number of false alarms and missed detections;
3. Multiple Object Tracking Precision (MOTP) reflects the 2D location precision of the tracking level; and
4. Multiple Object Tracking Accuracy (MOTA) is the tracking accuracy calculated from the number of false alarms, missed detections, and identity switches.

We repeated the experiment with six sets of parameters to observe tradeoffs in performance. Table 1 list the scores of the six runs. It can be seen that our system achieves reasonable results.

	Run1	Run2	Run3	Run4	Run5	Run6
MODP	0.5864	0.5804	0.5859	0.5792	0.5849	0.5837
MODA	0.4198	0.4198	0.4663	0.4552	0.4101	0.4654
MOTP	0.5861	0.5808	0.5851	0.5794	0.5840	0.5827
MOTA	0.4165	0.4162	0.4630	0.4516	0.4080	0.4632

Table 1. Final evaluation scores (the numbers in bold font are the best ones).

As MODA and MOTA metrics integrate missed detections and false alarms in a single number, it is difficult to see the tradeoff from these numbers. Instead, we use the fraction of ground-truth instances missed, and the number of false alarms per frame to draw an ROC curve, see Fig.8.

The videos of site 2 are more complicated than those of site 1 in terms of the variety of objects in the scene. Table 2 gives the scores of run 1 on these two sites. It can be seen that the performance on site 1 is much better than that on site 2. Site 2 is a street with a number of parking lots in front of some shopping area. Besides humans, there are many vehicles moving or parked on the street. The detection rate is lower because humans are often occluded by cars; while the number of false alarms increases because the blob tracker gives false positive on cars and the cars move slowly due to heavy traffic so that the speed based verification does not help much.

The speed of the system is about 0.2 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.

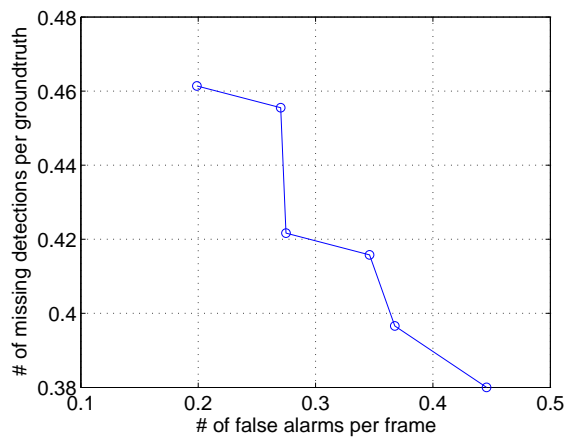


Fig. 8. ROC curve.

	MODP	MODA	MOTP	MOTA
Site 1	0.6167	0.6062	0.6172	0.6032
Site 2	0.5295	0.0884	0.5276	0.0085

Table 2. Scores on different sites.

4 Conclusion and Discussion

We applied a fully automatic multiple human tracking method to surveillance videos. The system has achieved reasonable performance on the test sequences. However, the performance does depend strongly on the complexity of the environment. Our future work will attempt to combine motion and shape cues in a stronger way to improve performance in more complex situations.

Our system does not run at real time. Some speedup can be obtained by code optimization and use of commodity parallel hardware. We can also obtain significant improvements by taking advantage of context in our algorithms.

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.

References

1. Fengjun Lv, Tao Zhao and Ramakant Nevatia. Camera Calibration from Video of a Walking Human, to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2006

2. 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
3. B. Wu, and R. Nevatia. Tracking of Multiple, Partially Occluded Humans based on Static Body Part Detection. 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. R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, M. Boonstra, and V. Korzhova. Performance Evaluation Protocol for Face, Person and Vehicle Detection & Tracking in Video Analysis and Content Extraction (VACE-II) CLEAR - Classification of Events, Activities and Relationships. [http://www.nist.gov/speech/tests/clear/2006/CLEAR06-R106-EvalDiscDoc/Data and Information/ClearEval_Protocol_v5.pdf](http://www.nist.gov/speech/tests/clear/2006/CLEAR06-R106-EvalDiscDoc/Data%20and%20Information/ClearEval_Protocol_v5.pdf)