Combining Edge and Color Features for Tracking Partially Occluded Humans

Mandar Dixit and K.S. Venkatesh

Computer Vision Lab.,
Department of Electrical Engineering,
Indian Institute of Technology, Kanpur
mandarddixit@gmail.com, venkats@iitk.ac.in

Abstract. We propose an efficient approach for tracking humans in presence of severe occlusions through a combination of edge and color features. We implement a part based tracking paradigm to localize, accurately, the head, torso and the legs of a human target in successive frames. The Non-parametric color probability density estimates of these parts of the target are used to track them independently using mean shift. A robust edge matching algorithm, then, validates and refines the mean shift estimate of each part. The part based implementation of mean shift along with the novel edge matching algorithm ensures a reliable tracking of humans in upright pose through severe scene as well as inter-object occlusions. We use the CAVIAR Data Set as well as our own IIT Kanpur test cases demonstrating varying levels of occlusion in daily life situations to evaluate our tracking method.

1 Introduction

Detection and Tracking of moving objects is central to many computer vision applications such as visual surveillance, activity recognition, and human computer interaction. The most commonly used methods for detection record the changes occurring in the scene. A statistical model of the scene background is learned and an intruding object is detected as a group of connected pixels not well represented by this model. [4] uses a single Gaussian, whereas, in [6], a mixture of Gaussians is used to represent the background and observe changes against it. Our system for tracking uses the foreground segmentation algorithm proposed in [7] and implemented by OpenCV [14] library functions. This method performs segmentation through Bayesian decisions on selected features representing static and moving scene elements.

Once detected, the object must be tracked in different frames using its signature. Features such as color, shape and texture may be used to establish correspondence between the occurrences of the same object in successive frames. Based on the nature of implementation, the tracking algorithms can be divided into feature based, contour based and region based categories. A Feature based tracker described in [8] uses corner points to track vehicles through traffic congestion. Point features, however, may not provide reliable means of tracking people through appearance changes. Contour based tracking demonstrated in [9] captures the shape information of the object. But these methods are generally slower than region based approaches. Among region

H. Zha, R.-i. Taniguchi, and S. Maybank (Eds.): ACCV 2009, Part II, LNCS 5995, pp. 140–149, 2010.

based tracking systems, the Kernel based mean shift tracker by Comaniciu et al [1] is well known. Given the target density estimate, the mean shift algorithm converges at the nearest mode of the point sample distribution represented by the test image. Although mean shift tracking provides accurate localization of an isolated object over short intervals of time, its performance degrades in the event of occlusion or change in the object scale or appearance. The color histogram used by the tracker changes substantially when the person being tracked turns about the vertical axis or is partially occluded by static or dynamic scene elements. A proposed solution to the problem of occlusion is the idea of coordinated tracking of multiple parts of the same target. Fragment based tracking (Frag-Track) proposed by Adam et al. [3] uses a template of fragments to track an agent through scale changes and partial occlusions. Frag-Track performs well in severe occlusions but the method uses a rigid template to describe a semi-rigid human body.

In this paper, we propose an approach to tracking that is a combination of region based and feature based paradigms. We use an efficient algorithm of matching local edges in conjunction with the Kernel based mean shift algorithm to obtain an accurate localization and track of humans in difficult scenarios. In order to use a more spatially descriptive appearance model for mean shift, we initialize three independent mean shift trackers for the head, torso and legs of the individual to be tracked. The part based approach ensures a better confidence through various levels of scene occlusion than the overall (single) mean shift. Following the mean shift cycles, the edge matching step validates and refines the mean shift estimates. Efficient tracking of people is achieved through coordinated mean shift and robust edge matching even in cases of severe occlusion, which is the key contribution of our research.

The rest of the paper is organized as follows. Section 2 summarizes some related research efforts to solve similar problems. Section 3 outlines our approach of edge-color tracking. Results of tracking in various complex situations are demonstrated in Section 4. In Section 5 we discuss the limitations of our method and future scope.

2 Related Work

Extensive research is being carried out in order to develop a system for tracking a moving target in a complex dynamic environment. One of the most well-known algorithms for object tracking is the kernel based mean shift proposed by Comaniciu et al. in [1]. The main advantages of Mean Shift algorithm are its speed of operation and accuracy of localizing moving targets. One of the drawbacks of this technique, however, is the lack of adaptability to scale changes. This problem has been addressed in detail in [5] and a scale invariant mean shift tracking procedure has been proposed as a possible solution. The author uses mean shift in spatial as well as scale dimensions to obtain an accurate localization and scale of the target. Zivkovik et al. propose a modifed mean shift procedure in [11] to include both scale and orientation changes by defning five degrees of freedom for the kernel.

Although kernel based mean shift is an effective region based algorithm for tracking isolated objects, its localization degrades in presence of occlusions and clutter. It is, therefore, not accurate enough for seamlessly following a moving target in complicated environments. Instead of using a single model for the entire object as in mean

shift tracking [1], a more descriptive part based or fragment based representation of the target is being preferred for better results in crowded scenes with frequent occlusions. Elgammal and Davis [2] propose that a person can be represented as a set of color regions located along the vertical axis. A person in upright pose is modeled as a collection of parts namely head, torso and legs, each having a separate color density representation. Along with color information, the spatial distribution of these parts is also included in the appearance model. In [3], the authors have used a fixed rectangular grid of patches with an intensity histogram of each patch to capture the spatial details along with the photometric information. The algorithm proceeds by finding the best matches of each fragment from the grid in the local neighborhood. Based on the similarity measures of the patches, a voting scheme is implemented to find an accurate estimate of the template center. To handle partial occlusions, Shakunaga et al. [10] propose an interesting technique that uses spatial information in a particle filtering framework. Their model represents a human using three ellipses, one each for head, torso and legs. Trained model based algorithms have also been developed for detection and tracking of humans in presence of persistent occlusions. Zhao and Nevatia [12] use a part based human model to solve a multiple hypothesis association problem. Efficient optimization in a joint hypothesis space is achieved using Markov Chain Monte Carlo method. Wu and Nevatia [13] propose a hierarchical part based model for detection and tracking of partially occluded people through trajectory estimation. Edgelet features from different parts of human body are used to train the model. The overlapping scores of detected part edges with the overall target segmentation are used to attribute part responses to a human hypothesis.

The proposed approach in this paper is different from the above mentioned ones in various respects. Our method uses the part based approach to track an object in an upright pose using a combination of both mean shift and edge tracking. Achieving better results by efficiently using a combination of features is the novelty of our approach. We implement mean shift tracking for head, torso and legs of the target independently. Our part based model is more flexible than in [3] as it does not impose a rigidness constraint on a semi-rigid human body. After convergence of each mean shift part tracker, an efficient edge matching algorithm validates and refines the estimate. Model based detection methods [12, 13] require extensive on-line learning. Instead we use background subtraction method to detect the target and learn mean shift kernels for each part. The Canny edge detection algorithm is used to extract strong edges of the target. The edge tracker uses information regarding curvatures and relative locations of stable object edges to match them over successive frames.

3 Proposed Approach

We propose a part based approach to track humans in an upright pose using a combination of edge and color information. The detected human silhouette is segmented to obtain the head, torso and legs and independent mean shift trackers are initialized for each part. An iterative edge matching step follows each mean shift cycle. The parts with high edge confidence indicate an accurate estimate. The mean shift trackers that diverge from their targets are thus separated from the faithful ones. Combined confidence of edge matching and color based mean shift helps in tracking the target

through severe occlusions with impressive localization. The steps involved in the proposed method are explained in detail in the following sections.

3.1 Appearance Model

As mentioned earlier, the lack of spatial information in the appearance model of a mean shift tracker can be remedied to some extent by employing a part based technique. To implement a part based variant of the traditional mean shift tracker, we segment the incoming background subtracted silhouette for parts namely, head, torso and legs. We search for the pronounced valleys in the horizontal projection of the foreground silhouette which mark the junctions between the parts we seek to locate. When the sensor optical axis is almost horizontal, the valley corresponding to head and torso junction can be located at a height 0.6 to 0.8 times the height of a reasonably good foreground blob, whereas the valley corresponding to torso and leg junction can be located at a height 0.3 to 0.5 times that of the blob. The blob segmentation method is demonstrated on a test image in Fig. 1. The non-parametric color probability densities for individual parts are then learned for their individual mean shift trackers.

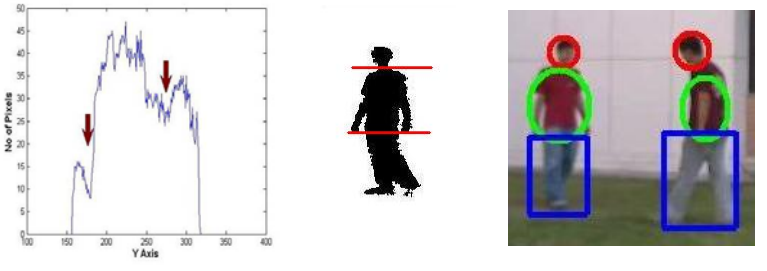

Fig. 1. Segmentation of blob along valleys in vertical projection

Although a part based appearance model ensures that at least one of the trackers follows the target accurately, the deviant behavior of a mean shift tracker in clutter or due to occlusion necessitates a method to verify the credibility of the individual part mean shift estimates. As a result, we supplement the color information of the appearance model with the local edge information which defines both shape and texture of a human target. Strong edges in the region of the foreground blob are obtained from the image using Canny edge detection algorithm implemented in OpenCV library functions. These learnt edges would then be matched with those extracted in following frames using their positions and curvature features. The edges obtained from a human image are relatively less stable when compared to those of rigid objects. Nevertheless, the change in their curvature and location is gradual enough to enable matching over a short interval of frames after which the template has to be reinitialized.

3.2 Mean Shift Part Tracking

As mentioned earlier, we use independent mean shift trackers to follow the head, torso and legs of a person. We use the Epanechnikov kernel to find the density

estimate of RGB color values of the segmented pixels. The probability of a color u as expressed by the kernel density function can be written as:

$$P_M(u) = C \sum k(\|x_i / h\|)\delta[b(x_i) - u], \quad u = 1, 2, \ldots, M \tag{1}$$

Where M denotes the number of histogram bins, x_i denotes the pixel location and k(.) denotes the profile function of the kernel. For details of mean shift tracking algorithm the readers are referred to [1].

The main drawback of mean shift tracking is the drift of the tracker due to occlusion and clutter. Fig. 2 shows the effect of scene occlusions on the mean shift algorithm. Such divergence of the mean shift tracker may cause complete track loss. Our methods prevent the degradation in performance due to occlusion through the use of a part based model. The trackers corresponding to un-occluded portions of the target maintain proper track throughout and thus, prevent incorrect localization as would happen in the case of traditional mean shift tracking.

Scale Handling

The Mean shift tracking algorithm does not account for scale changes of the target. The techniques of updating target scale proposed by [5, 11] are interesting but require heavy computations. We, instead, use a simple method to handle scale changes in our algorithm. The number of foreground pixels is a useful heuristic indicating change in the size of a target. Although, foreground blobs are not always reliable indicators of the shape and size of the object, we can identify scale changes since they cause blob size to alter gradually. We set thresholds on percentage changes in number of foreground pixels of the blob and reinitialize the mean shift kernel when the change stays within this threshold value. A blob change that exceeds the assigned threshold indicates either occlusion or improper segmentation, in the event of which, the mean shift kernel is kept unchanged. This simplified method provides satisfactory handling of the scale variation problem.

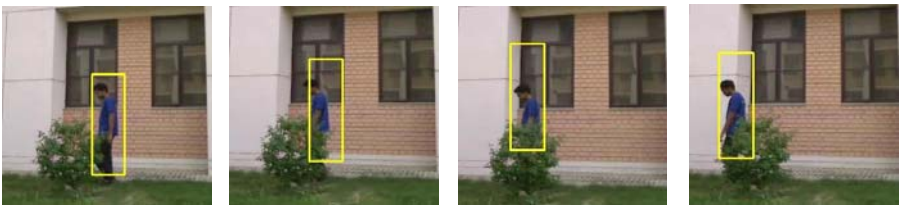

Fig. 2. The sequence of images demonstrates the divergence of mean shift tracker from the target in presence of scene occlusion

3.3 Edge Matching

We use a robust edge matching algorithm in conjunction with the part based mean shift to improve tracking performance in difficult and crowded situations. An edge can be identified by its location on the object and its curvature. We use both these

features to match the extracted edges with those in the learnt template. To capture the curvatures of an edge along its length, we model it using straight line segments of fixed lengths. The orientations of these straight line segments are recorded as an estimate of the edge curvatures. Smaller the length of these segments, greater is the accuracy of the estimate.

Although edges are a reliable feature for tracking, the edges of a moving human target change over time as against the edges of a rigid object. As a result, the edge template needs to be reinitialized when necessary. We use percentage change in the blob size as an indicator of changing scale and orientation (about the vertical axis) to reinitialize both the mean shift kernel and the edge template, by setting up an upper as well as lower threshold. The former is required to prevent spurious updates possibly caused by segmentation failures or occlusions.

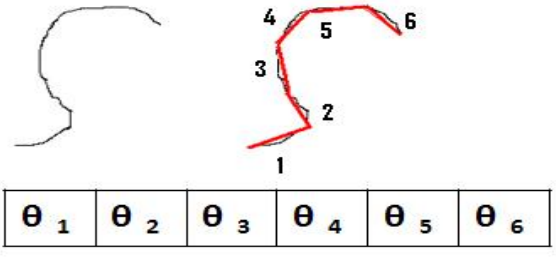

Fig. 3. A schematic showing edge approximation with straight line segments (in red). The vector $\{\theta_i\}$ indicates the orientations of these straight line segments (with respect to horizontal).

3.3.1 Algorithm for Edge Matching

An Edge e can be represented by a vector θ of orientations of the straight line segments approximating the edge, as shown in Fig. 3, and a representative point p located on it. Suppose we wish to calculate the matching score of an extracted edge e_m with a learnt edge e_t from the template. We denote their orientation vectors as θ_m and θ_t of lengths l_m and l_t respectively. Edges e_m and e_t may be same or different or one may be a part of the other. To verify a match between the two, we must locate the set of points common to the two edges. The straight line segments approximating the two edges that lie in this region of match have almost the same orientations. That is, one can ascertain a match between two edges by locating matching sections in the two orientation vectors. This can be depicted as sliding the orientation vector of one edge over another and matching the directions of overlapping sections. If the two edges match, the mean absolute orientation difference between the overlapping sections of their vectors would be negligible or zero. The problem of curvature matching between edges can, thus, be formulated as one of finding the minimum mean absolute orientation difference between their vectors and comparing it with a threshold.

$$\Theta(n) = \sum_k | \theta_t(k) - \theta_m(n+k) | / \Omega(n) \quad , n = -l_m + 1, -l_m + 2, \ldots , l_t - 1 \tag{2}$$

The term $\Omega(n)$ represents the overlap between two vectors θ_t and θ_m, corresponding to the value of index n, which indicates their relative positional displacements. The value of $\Omega(n)$ could be calculated as

$$\Omega(n) = \min(l_t, l_m) - n \tag{3}$$

When the match between the overlapping sections of the edges is perfect, the value of mean absolute orientation difference $\Theta(n)$ diminishes. This residual value is denoted as $\Theta_{m,t}$, the minimum difference between the curvatures of two edges e_t and e_m. The corresponding overlap $\Omega_{m,t}$ is the overlap of best curvature match between them.

$$\min_n \Theta(n) = \Theta(r) = \Theta_{m,t} \tag{4}$$

$$\Omega(r) = \Omega_{m,t} \tag{5}$$

Since different edges may have the same curvature, taking their location into consideration is of prime importance. After matching the edge curvatures, we find the midpoints of the overlapping parts of both edges. Suppose *p* and *q* denote the midpoints of overlapping sections of the edges. (Note: the coordinates of p indicate location of the point from the top-left corner of the mean shift window, whereas coordinates of q on the learnt edge indicate its location from the top-left corner of the template window. The proposed idea is that if a mean shift tracker maintains accurate track of an object, the relative locations of its edges on the target remain nearly the same over successive frames). The proximity of edges is indicated by the Euclidean distance between these two points. We modify the minimum difference metric of the edges to include this distance information and the value of edge overlap $\Omega_{m,t}$ to prevent false matches.

$$\gamma_{m,t} = \min_n \Theta(n) \times d(p, q) / \Omega_{m,t} \tag{6}$$

where $d(p, q)$ is the distance between the midpoints *p* and *q* of the overlapping parts of the edges. This, then, is the procedure and the metric devised to match two edges.

In case of a human target, after every mean shift cycle, the local edges present in every part mean shift window are matched with the edges of their respective edge templates. That is, we need to find pairs of edges e_m from the current frame and e_t from the learnt template, corresponding to a particular part, that exhibit a match with the value of their metric $\gamma_{m,t}$ lying within a predefined threshold T. Every such pair of matching edges would lie at a certain Euclidean distance $d(p, q)$ with respect to each other. If we reposition the mean shift window such that this distance between the query edge and the template edge diminishes, they would show a better match (lower value of $\gamma_{m,t}$). In other words, if the average of Euclidean distances $d(p, q)$ for all the pairs of matching edges is used to alter the mean shift window location, an overall improvement in the edge matching result would accrue. Such adjustment in the mean shift window may also result in newer pairs of matching edges. Hence, the process of edge matching (estimate validation) and window repositioning (estimate refinement) are carried out iteratively until a convergence is reached. The overall edge matching confidence in each match cycle is calculated as follows:

$$C = \sum_{m,t} \Omega_{m,t} \times \log (T/\gamma_{m,t}) \quad (7)$$

The sum is over all pairs of matching edges (m, t). Fig. 4 shows the results of edge matching algorithm on some human test images.

4(a) 4(b) 4(c) 4(d)

Fig. 4. (a) and (c) represent the learnt templates of head and torso edges respectively. (b) and (d) show the corresponding matched edges.

3.4 Part Assignment and Target Localization

Following the part mean shift cycle, the edges present in the mean shift windows are matched with the part edges present in their respective learnt edge templates. Parts being accurately tracked would exhibit a high Bhattacharya matching as well as edge matching confidence. Based on the dimensions and locations of such part trackers, an elliptical bound is derived to mark the target. In each frame, its dimensions and positions are updated based on the part tracking responses. If one of the trackers deviates due to clutter in the background or occlusion, it ceases to show a good edge matching confidence (C in eq. 7). If the confidence falls below a set threshold, we ignore the tracker completely and update the target marker position and dimensions solely based on the remaining faithful trackers. If both the torso and the leg trackers of the human target diverge, only the position of the ellipse is updated according to that of the head tracker and the dimensions are kept unchanged.

4 Results

The proposed algorithm was used to track humans through various events of occlusions as seen in Fig. 4. We use the standard CAVIAR Dataset (*http://homepages.inf.ed.ac.uk/rbf/CAVIAR/*) to evaluate our tracking algorithm. The first and second sequences shot in a corridor of a mall show the satisfactory performance of our algorithm in presence of almost 70-80% occlusion (head and a small part of torso visible). We also test our algorithm on our IIT Kanpur dataset demonstrating various scene occlusion scenarios. The third and fourth sequences show a person being occluded by a shrub and parked two-wheelers respectively. Our combined Edge-Color tracker (ECT) maintains accurate track throughout the event of occlusion in both sequences.

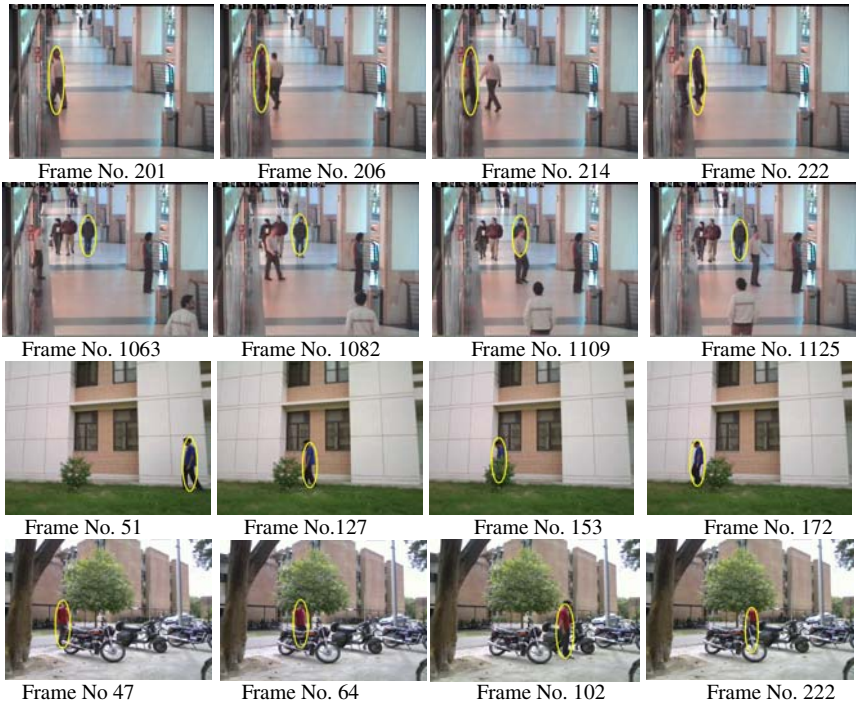

Fig. 5. Performance of ECT on CAVIAR test cases and IIT Kanpur Dataset

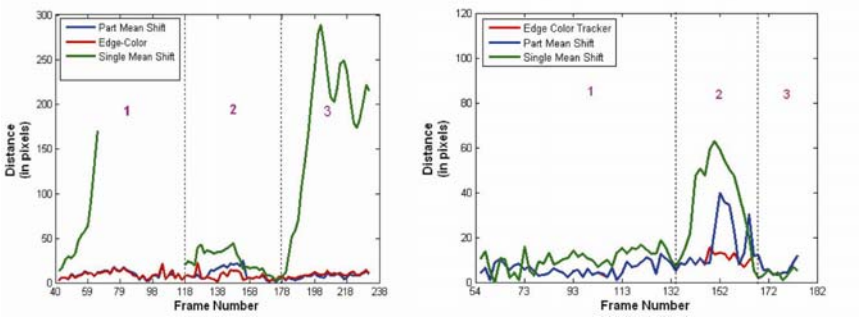

Fig. 6. Both the plots represent the difference (in pixels) between the tracker localizations and manually determined ground truth. Plot in Red indicates edge color tracker localization. Single mean shift tracker localization is shown in Green whereas a simple part mean shift performance is shown in Blue. The plot on the left shows the performance of both trackers in "parking space" sequence while the plot on right shows their performance on "shrub data". The discontinuity in the green line on the left plot shows that the mean shift tracker was lost and needed to be re-initialized while the edge color tracker kept good track.

Fig. 6 shows plots of localization errors of trackers with respect to manually marked ground truth for shrub and two-wheeler parking sequences of Fig. 5. A progressive improvement is seen from a single kernel mean shift (shown in green) through a simple part based mean shift tracker (blue) to the Edge Color Tracker (red). The edge matching algorithm provides means to verify the credibility of part mean shift trackers. As a result the localization of an ECT is more reliable than just a part mean shift tracker in presence of clutter or occlusion.

5 Conclusion and Future Work

In this paper, we have proposed a simple yet highly effective technique for tracking partially occluded humans in a standing/walking position using a combination of edge and color features. One of the drawbacks of our method is its failure to update scale during persistent occlusions. The scale adaptation approach we use requires the complete foreground blob of the object. Hence, scale cannot be updated during the event of occlusion. Incorporating scale invariance in a more reliable manner in the proposed tracking framework would be the focus of our future efforts.

References

[1] Comaniciu, D., Ramesh, V., Meer, P.: Kernel-Based Object Tracking. IEEE Trans. PAMI 25(5) (2003)
[2] Elgammal, A.M., Davis, L.S.: Probabilistic Framework for Segmenting People under Occlusion. In: Proc. ICCV (2001)
[3] Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments based tracking using the integral histogram. In: CVPR 2006, pp. 798–805 (2006)
[4] Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-time tracking of the human body. IEEE Trans. on PAMI 19(7), 780–785 (1997)
[5] Collins, R.T.: Mean-shift blob tracking through scale space. In: IEEE CVPR, vol. 2, pp. 234–240 (2003)
[6] Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: Proceedings CVPR, pp. 246–252 (1999)
[7] Li, L., Huang, W., Gu, I.Y.H., Tian, Q.: Foreground object detection from Videos containing complex background. In: ACM MM 2003 (2003)
[8] Beymer, D., McLauchlan, P., Coifman, B., Malik, J.: A real-time computer vision system for measuring traffic parameters. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR (1997)
[9] Freedman, D., Zhang: Active contours for tracking distributions. IEEE Trans. Image Proc. 13(4), 518–527 (2004)
[10] Satake, J., Shakunaga, T.: Multiple target tracking by appearance-based condensation tracker using structure information. In: ICPR (2004)
[11] Zivkovic, Z., Krose, B.: An em-like algorithm for color histogram-based object tracking. In: IEEE CVPR (2004)
[12] Zhao, T., Nevatia, R.: Tracking multiple humans in crowded environment. In: CVPR, vol. II, pp. 406–413 (2004a)
[13] Wu, B., Nevatia, R.: Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors. In: IJCV 2007 (2007)
[14] `http://sourceforge.net/projects/opencvlibrary/`