Local Segmentation for Pedestrian Tracking in Dense Crowds

Clement Creusot

Toshiba RDC, Kawasaki, Japan,
clementcreusot@gmail.com,
http://clementcreusot.com/pedestrian

Abstract. People tracking in dense crowds is challenging due to the high levels of inter-pedestrian occlusions occurring continuously. After each successive occlusion, the surface of the tracked object that has never been hidden reduces. If not corrected, this shrinking problem eventually causes the system to stop as the area to track become too small. In this paper we investigate how hidden parts of one target object can be recovered after occlusions and propose challenging data to evaluate such segmentation-tracking technique in dense crowds. The segmentation/tracking problem is particularly difficult to solve for non-rigid objects. Here, we focus on pedestrians whose limbs and lower body parts often get occluded in crowded scene. We first investigate the unmet challenges of pedestrian tracking in crowds and propose a challenging video to evaluate segmentation-tracking robustness to inter-pedestrian occlusions. We then detail a fast segmentation-based method to overcome some aspects of the tracking-under-occlusion problem. We finally compare our results with two existing tracking methods.

1 Introduction

Tracking a single object/person in a video is a trivial task for any human operator. Machine vision systems are able, to some extent, to execute such a task. One of the most successful approaches used by many researchers has been to combine frequent object-detection with object-tracking algorithms. While these systems are not as robust as human perception, they now achieve very high results on video of sparse pedestrian scene (*e.g.* [1–3]). The success of these approaches is mainly due to the frequent human body or head detections which are very fast and fairly robust in non-occluded scene. The main limit of such techniques occurs when the detector is no-longer able to seed an initial solution to the tracking algorithm. This happens when pedestrians start to be occluded by other objects, which most of the time are also pedestrians. At this point, at least two research directions are possible: First, to develop even more sophisticated detectors that will reliably detect body parts under occlusion (*e.g.* Poselets [4]). This is a sound idea as humans are able to segment pedestrians' body parts from still images. Second, to develop trackers that do not rely as strongly on detection.

In this paper we focus on the second approach. We believe that robust tracking cannot be done while simply considering rectangular tracking windows. Indeed, a large proportion of the pixel in a tracking window will not belong to the tracked object (See Figure 1a). An efficient tracking system should know for every frame the set of pixels belonging to the object it is tracking. This segmentation problem is particularly difficult to solve in crowded scenes. One problem with the tracking/segmentation approach is that the area to track has a tendency to shrink over time as show in Figure 1b. In this paper we investigate simple prior-free methods to try to recover these parts.

(a) (b) (c)

Fig. 1. (a) Rectangle Window vs. Segmentation: here the pedestrian represents only a quarter of the rectangle selection area. (b) Point-tracking "shrinking" problem cause by occlusions. (c) Example of post-occlusion lower limbs recovery using our system.

The target application for this work is automatic or assisted surveillance. Requirements for automatic surveillance include people detection, tracking, re-identification, action recognition and ultimately suspicious behavior detection. While we focus on tracking in this paper, the automatic segmentation of pedestrians can be beneficial to all these different tasks. In non-automatic scenarios the people to track (suspicious individuals) are selected by surveillance personal. In practice the number of suspicious people will be far lower than the number of people in the scene. Tracking everybody approximately seems therefore less important than following one person in a robust way. Following a single individual also allows for computation to be local and therefore more efficient.

This paper is at the intersection of video segmentation and people tracking. The main difference with other tracking papers is that we focus on different performance measures (time before failure without detection and segmentation errors) as well as different challenges (heavily occluded pedestrians in crowds). The technical contributions of this paper are a novel heuristic method using segment sampling to re-grow missing regions after occlusions; a dense ground-truth dataset which is ideal for occlusion-robustness evaluations of pedestrian tracking method; and a fine-grain evaluation of segmentation errors for segmentation-tracking.

In the following sections, a brief literature review is given before presenting our proposed evaluation benchmark. The proposed method is detailed in sec-

tion 4. The results are presented in the following section. A discussion about how to approach pedestrian tracking in crowds in given in section 6.

2 Related Work

Most of the venues that require automatic assistance for video surveillance are also the places that show the highest level of pedestrian traffic and dense crowds (train stations, airports, department stores, stadiums and so on). Therefore, systems that can not deal with dense crowds are likely to be unusable in the very places they are needed the most. This has been acknowledged many times in the literature, but solutions for such problem are still scarce. Indeed a majority of pedestrian tracking methods only show results on videos of sparse set of people presenting limited and sparse cases of inter-pedestrian occlusions. In [5], first approaches have been tested to extend pedestrian detection-based tracking to crowds. In [6], a Bayesian clustering approach is used to group moving keypoints into individual pedestrian clusters. This model-free method is mainly based on motion information and show interesting results. However the density of point detection is usually low and, when apply to crowds, the system ends up tracking mainly the heads of people. Recently, [7] proposed a dense segmentation method to increase the density of tracking points. They cluster points motion into small segments that are then merged into larger segments using a motion invariant (geodesic) distance between segments. The final label attribution is done via an offline global optimization. In [8], a method to conciliate pedestrian detections and point trajectories was proposed. The tracking hypothesis based on detection windows (based on poselets [4]) and local optical flow motions are combined to get a more robust overall tracking.

In terms of data, there is no obvious benchmark to evaluate robustness to occlusion in crowds. Often crowds video present uniform motions. Indeed, corridors, sidewalk and zebra crossing scenes are defined along a single axis which can be used unintentionally by the researchers to smooth the trajectories. We consider that scene showing mutli-directional free motions are more appropriate to evaluate tracking algorithms. In [9], a dataset of crowded scenes was proposed. While most of the scene show crowds captured from long range which are not suitable for individual tracking, some of the videos show intermediate camera views with complex non-uniform crowd motion that are ideal to challenge tracking systems. In this paper, the density of the crowd refers to the number of people per surface area on the scene floor, not the number of people per pixel area in the image.

3 Problem Analysis

When tracking people in dense crowd the main problem is often to determine what pixels or patches of the input to track. A good description of this problem is given in [10]. If a rectangular window patch is tracked, either too few or too much data is used depending on the window size. A window that encompass

the whole pedestrian is indeed polluted by background, occludees and occluders pixels. In order to track only the part of that rectangle corresponding to the object it is necessary to segment the target and track only features associated with these pixels.

In this paper, we assume the subject detection at the initial frame to be given as input to the system. For real applications this information can be provided by an automatic pedestrian detector (before any severe occlusion happens) or in another scenario by a human operator first selecting a suspicious individual to track on a security screen. The semi-supervise tracking scenario might sound counter-intuitive: if you have an operator to detect suspicious individuals why not ask him to do the tracking? The reason is simple: the operator would not be able to perform any other tasks (see unintentional blindness [11]) whether it is tracking someone else, answer the phone or detect other suspicious people. Besides, tracking is a very tedious task that requires constant attention and would grow tired any operator very quickly. Monitoring several automatically-tracked people is more easy than tracking one person manually. Our particular objective in this paper is to track a single target individual for as long as possible without intermediary automatic detections.

a b c

Fig. 2. Proposed benchmark data for segmentation-tracking evaluation in dense crowd: (a) Crowd scene video (879-38_1) from the UCF Crowd Dataset. (b) Example of manual segmentation for 10 pedestrians at frame 100. (c) 3D visualization of our temporally dense segmentation ground-truth for 300 frames.

3.1 Data

The main objective of our system is to improve robustness to occlusions observed in dense crowds. We choose one very challenging video (879-38_1.mov) from the UCF Crowds dataset [9] for our main evaluation. This video scene (see Figure 2a) is a scripted set up of numerous extras coming in and out of the screen in a dense crossing area. The average number of pedestrians per frame is around 30. The total length of the original video is 1285 frames (51.4 seconds). Sidewalks, corridors and two-ways zebra crossing scenes are limiting in that they only show

two main directions for pedestrian motion. Here the actors cross each other in all possible directions making it an ideal benchmark to test against body occlusions. Most of the actors wear suits (often dark) which are difficult to segment from each other when overlapping. The video is relatively low resolution (480×360) which make texture-based method less efficient. Despite being directed, the scene is a good representation of difficult cases that can occur in busy train stations or airports halls. The video present lot of the challenges observed in real life situation which makes it a good support for evaluation.

We consider pedestrians within the first 300 frames (12 seconds) of the videos. Since it was not possible to segment all pedestrians, we selected 10 for which the ground-truth has been manually defined for every frame (see Figure 2b). The manual ground truth we have generated is freely available for download on the author webpage [1]. To our knowledge it is the first temporally dense real video segmentation ground-truth available for a crowded scene.

4 Proposed Approach

Our first intuition to solve this problem is that very little information is required to track pedestrians.

When looking at a video of edge-segmented frames, a human can track the pedestrians without difficulty despite the lack of texture. The minimal amount of information for pedestrian motion detection/tracking have been studied for a long time for skeletons [12] and more recently for blob-like patterns in "emerging" images [13]. These studies strongly indicate that very little input information is required for human to track moving shapes. It is possible that the human brain use some sort of shape prior to perform such task. What is important to note here is that a video presenting only edges and motion contains enough information to solve the pedestrian segmentation/tracking problem. The question remaining is what general rules are needed to infer the solution from such input: 3D scene logic, continuity of motion and inertia, potential shape priors?

For this experiment we deliberately discard the direct use of texture information. The input images are used solely to generate segmentation and edge data, as well as to compute pixel displacement between frames.

4.1 Workflow

The input of our system is made of a video of a scene and an initial segmentation of one pedestrian. Our aim is to output for each frame of the video the propagated mask of this pedestrian despite all the occurring occlusions. Our process (see Figure 3) follows this following steps.

Mask Propagation (Push) - A first approximation of the pedestrian segmentation is computed by propagating points belonging to the previous frame mask

[1] http://www.clementcreusot.com/pedestrian/

Fig. 3. Process workflow.

to the current frame. The perceived motion of small local neighborhoods around these points is used as displacement. The mask at $t-1$ is sampled using a sparse grid of 3x3 pixels. These points are pushed to the current frame using the standard Lucas-Kanade pyramidal optical flow [14]. We eliminate less robust elements by performing a back propagation of the keypoints using the reverse optical flow. We select only the initial keypoints which verify the following (unit in pixels):

$$Dist_{Eucl}(k, \tau^{-1}(\tau(k))) < 10 \tag{1}$$

where τ is the optical flow translation function between frame $t-1$ and t and τ^{-1} the translation function between frame t and $t-1$.

Point Filtering - The optical flow is likely to give motion in different direction for the same object. In order to eliminate as much noise as possible we filter the points obtained using their associated translation direction. The mean direction for the last 5 frames are kept in a queue and used to filter incoherent displacements. The directions are clustered using their angle differences to the past mean and inliers within 0.5 radian of the main cluster centroid are used to compute the new mean. Points that are selected with this method are relatively sparse, especially in region presenting uniform texture. However they show coherent motion direction (See Figure 3).

Superpixel Expansion (Grow) - The propagated points provided by the previous method are sparse. If only those points were tracked, the system will fail after just a few frames. In this paper we investigate two different segmentation approaches: a simple Superpixel approach [15] and a hybrid method between edge-detection and superpixels. The segmentation is grown by merging the interior of all segments/superpixels which contain more than 5% of keypoints. If the superpixel usually follow the edges of objects, it is not always the case and this segmentation will have a tendency to leak outside the boundary of the object. Besides, in some case the occluder and occludee have exactly the same texture where the occlusion happens (*E.g.* sportwears during marathon, business-men dark suits, and so on). The masks retrieved by this method can also present some holes corresponding to regions where very few points were pushed.

Our system is object independent except for one parameter. When merging superpixels together we consider that the cost of merging horizontally is twice

the cost of merging vertically. This translate the idea that a person shape is more likely to be vertical than horizontal. The cost is computed as $d = 2dx + dy$. All superpixel containing target points are ordered according to the position of their centroid relative to the current shape center. Areas are merged in this order until the set of superpixel candidates is empty or the maximum area is reached (the area of the object at the initial frame).

4.2 Implementation Details

The program is written in C++ using the OpenCV library and runs in around 0.09 second per pedestrian per frame. The hybrid segmentation is performed by detecting canny edges from which a first segmentation is produced using a ball filling of 7 pixels to close gaps between edges. Regions not detected by this method are segmented with a maximum of 50 superpixels [15]. In this article we only consider fast segmentation techniques that can be used for almost real-time applications. Level-set and graph-cut segmentation techniques are not discussed in this paper. Flood-fill segmentation methods have been tested but showed very low performances and are not presented here.

5 Experiments

Here we evaluate how this approach compares with two existing tracking systems and analyze the segmentation errors using our manual ground-truth.

5.1 Results

Qualitative results of our method are best seen in video. The videos presented on the author webpage [2] show the tracking results for the three tested methods as well as the segmentation error over time.

Time before failure - One of the most important criteria for this method is to be able to follow the object for the longest period of time without any intermediary pedestrian detection. For each pedestrian we give the number of frames successfully tracked before the method fails. We consider that the tracking has failed if the recovered area represents less than 10% of the ground-truth mask for two consecutive frames or if the label changes.

In Figure 4a we see the influence of the segmentation method used. The techniques marked as SP-X are using superpixel segmentations with at most X cells in the local tracking area. The hybrid method is the one using edges (see Section 4.2). In Figure 4b we compare with [8] (Two-Granularity Tracking) and [16] (Compressive Tracking). Most of the tracking techniques we have tried completely failed on this crowd video as they rely heavily on detection. The best achieving existing tracking paper we have found on this dataset is [8] which

[2] http://www.clementcreusot.com/pedestrian/

was designed to compensate for the lack of detection and be more robust to occlusions. Here we plot the results from the first detection of the person by their system until the tracking is lost or the label changed. Among the 10 people we are monitoring, the longest period before failure with [8] is 70 frames (for id 5) and this case occur before large occlusions (see Figure 5). Most of the time the tracking is lost or the label is switched. To their defense, we do not believe that such dense crowd was considered when designing the method. Tracking people in dense crowd seems to raise a different set of problems altogether.

The results presented here differ a lot from what is usually presented in tracking papers. What is the point of a tracking that fails after 100 frames? This is because we evaluate visual tracking *methods/modules* rather than tracking *systems*. The aim here is to track for the longest possible time under occlusion *without* intermediary detections. By forbidding intermediary detections we can observe more closely the flaws of the visual tracking techniques.

It has to be noted that the compressive tracking method [16] is extremely sensitive to the initial detection window. If the initial bounding box is too large or too small the system fails after just a few frames. We therefore consider this method to be less robust than [8]. Note that in Figure 4, the results given for [16] are the best (max) results obtained over several runs introducing random noise on the initial window (not the mean).

In Figure 5 the level of occlusion of each pedestrian during the sequence is plotted. As expected the tracking failures are often correlated with pedestrian occlusions (peaks in the graphs). The person the most difficult to track in this experiment is pedestrian number 4. One possible explanation is that this person is already heavily occluded at initialization (50% occluded at frame 0).

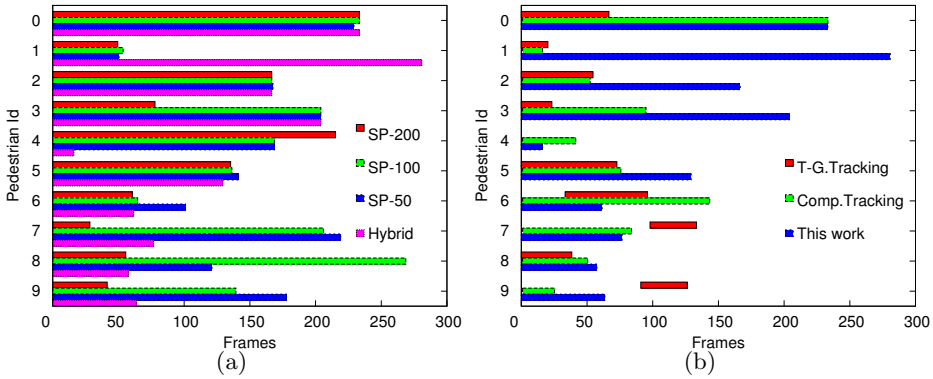


Fig. 4. (a) Results using different segmentation methods. (b) Comparison of the proposed method with [8] and [16]. Note that the detector used in [8] does not always give candidates for the first frame of the sequence. We plot the results from the first detection to the first failure.

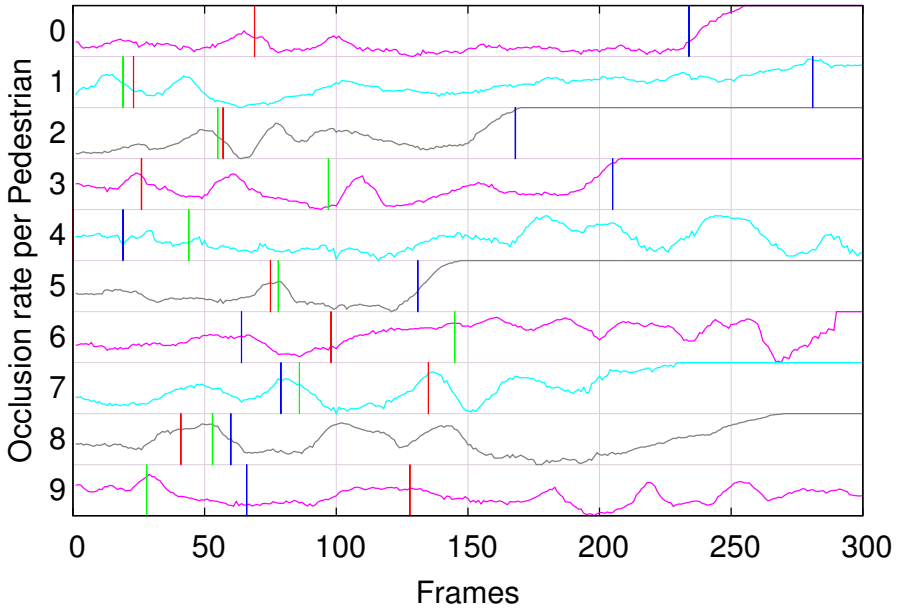


Fig. 5. Curves representing the ground-truth percentage of occlusions for each pedestrian in the sequence. A plateau at 100% of occlusion is reached when the pedestrian is off screen. The vertical lines represent the first-failure of the Two-Granularity tracking (red), Compressive tracking (green) and our segmentation tracking (blue).

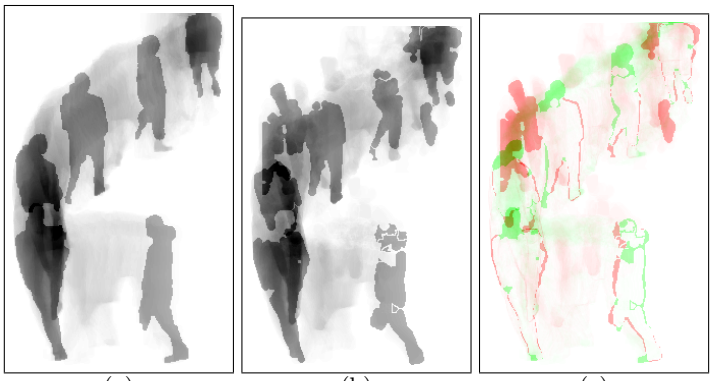


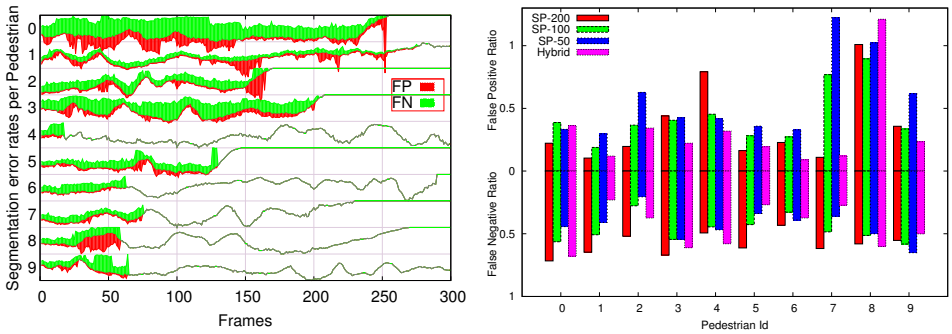
Fig. 6. Qualitative results for pedestrian 1: (a) Ground-truth time-lapse mask. (b) Computed time-lapse mask. (c) Differences, missing area are shown in green (false negative), extra area in red (false positive).

Segmentation Errors - For each pedestrian and for each frame the difference between the recovered mask M_t and the ground-truth G_t can be computed in terms of false positive pixels (extra pixel that were not present in the ground-truth) and false negative pixels (missing pixels). We normalize these values using the total number of pixels in the ground truth mask in that frame.

$$FP_t = \frac{|M_t \setminus G_t|}{|G_t|} \quad \text{and} \quad FN_t = \frac{|G_t \setminus M_t|}{|G_t|} \tag{2}$$

In Figure 7, the hybrid segmentation is compared to standard superpixel approaches. An interesting point is that the pedestrian who is tracked for the longest period (pedestrian 1) is also the one that shows the lowest segmentation errors. However the correlation between the two metrics (segmentation error and time of tracking) is not straight-forward. While the hybrid segmentation gives better accuracy, it is in average less robust in time than simple superpixels (at least for pedestrian above id 4, that are usually further away from the camera and more occluded on the starting frame).

Please note that these results are raw data. We did not perform any temporal smoothing to detect and correct segmentation errors. Indeed this would make it more difficult to compare the underlying segmentation techniques. Visualization of the segmentation errors can be viewed for one pedestrian as a time-lapse in Figure 6 and for all pedestrians as a video in the supplemental materials.



(a) (b)

Fig. 7. (a) False Positive (FP) and False Negative (FN) errors over time for the hybrid segmentation. (b) Averaged FP and FN ratios for the different segmentations used in the growing stage.

5.2 Limitations

The proposed approach is very simple in its structure and relatively fast (0.09s per pedestrian per frame). However it focuses only on one segmentation/tracking

issue. This process is thought to become essential when pedestrian detectors start to fail due to large occlusions and when visual tracking is required for many consecutive frames without re-detection. Our current approach is not using any object related prior (except for the merging cost indicating that the shape is mostly vertical). Hence, we can not segment touching object that move in the same direction for a long time. If the segmentation of the first object leak to the second object, the second object will be considered a part of the first object. Simple approaches can help deal with this situation. Co-segmenting objects within a close neighborhood can help resolve that kind of conflict where several objects claim a common set of pixels.

6 Discussion

In this article we presented a very challenging video and ground-truth for pedestrian tracking and investigate how previously hidden body parts can be recovered implicitly using local segmentation techniques. We analyzed these techniques after noticing that all conventional tracking methods fail on dense crowd videos, and posed the hypothesis that edge and motion are sufficient to perform pedestrian tracking.

It goes without saying that pedestrian tracking by visual tracking (including tracking by segmentation) and by detection are complementary. Any real-world applications will necessarily use a combination of those within a system and might have to deal explicitly with non-moving objects, long-term total occlusions and re-identification. We believe that more robust visual tracking and segmentation can improve tracking systems performance dramatically. Especially when frequent occlusions make pedestrian detection almost impossible.

One of the main conclusion of our experiments is that methods based on motion and edges can do much of the work required for a tracking-segmentation module. It provides a simple and relatively fast way of following and segmenting objects under occlusions for a longer period of time than "black-box" rectangular-window tracking techniques.

What we discovered while looking at difficult crowd datasets is that most of the techniques used to improve tracking on sparse pedestrian scenes do not extend well to crowded environments. This is somewhat worrying. Is it possible that the current tracking systems over-fit simple and sparse pedestrian scenes? Would it be possible to perform tracking on dense crowd data by building upon current state-of-the-art techniques or it is necessary to deconstruct them and start over with this new type of data in mind?

Our opinion is that a smooth transition from the techniques working well on sparse scenes to techniques working well on crowded scenes is possible if and only if fast occlusion-robust pedestrian detectors can be designed (*e.g.* fast poselets). If they cannot, we believe that the foundation for real-time pedestrian tracking in dense crowd will be very different from the current state-of-the-art tracking methods and might rely more on video segmentation techniques. Future work will investigate the use of priors (pedestrian shape statistical model, *e.g.* [17]) as

well as spatio-temporal segmentation (*e.g.* [18]) to constrain the growing phase of the algorithm.

References

1. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: CVPR. (June 2011) 3457–3464
2. Yang, B., Nevatia, R.: Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In: CVPR. (2012) 1918–1925
3. Andriyenko, A., Schindler, K., Roth, S.: Discrete-continuous optimization for multi-target tracking. In: CVPR. (2012)
4. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: ECCV 2010. Volume 6316. Springer Berlin Heidelberg (2010) 168–181
5. Zhao, T., Nevatia, R.: Tracking multiple humans in crowded environment. In: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. Volume 2. (2004) 406–413 Vol.2
6. Brostow, G.J., Cipolla, R.: Unsupervised bayesian detection of independent motion in crowds. In: IEEE Computer Vision and Pattern Recognition. (2006) 594–601
7. Iwasaki, M., Komoto, A., Nobori, K.: Dense motion segmentation of articulated objects in crowds. In: ICPR. (2012) 861–865
8. Fragkiadaki, K., Zhang, W., Zhang, G., Shi, J.: Two-granularity tracking: Mediating trajectory and detection graphs for tracking under occlusions. In: ECCV. (2012) 552–565
9. Ali, S., Shah, M.: A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on. (2007) 1–6
10. Aeschliman, C., Park, J., Kak, A.: A probabilistic framework for joint segmentation and tracking. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. (2010) 1371–1378
11. Simons, D.J., Chabris, C.F.: Gorillas in our midst: sustained inattentional blindness for dynamic events. Perception **28** (1999) 1059–1074
12. Johansson, G.: Visual perception of biological motion and a model for its analysis. Perception & Psychophysics **14**(2) (1973) 201–211
13. Mitra, N.J., Chu, H.K., Lee, T.Y., Wolf, L., Yeshurun, H., Cohen-Or, D.: Emerging images. ACM Transactions on Graphics **28**(5) (2009) 163:1–163:8
14. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI). (1981) 674–679
15. Ren, C.Y., Reid, I.: gslic: a real-time implementation of slic superpixel segmentation. Technical report, University of Oxford, Department of Engineering Science (2011)
16. Zhang, K., Zhang, L., Yang, M.H.: Real-time compressive tracking. In: Proceedings of the 12th European conference on Computer Vision - Volume Part III. ECCV'12 (2012) 864–877
17. Baumberg, A., Hogg, D.: Generating spatiotemporal models from examples. Image and Vision Computing **14**(8) (1996) 525 – 532
18. Apostoloff, N., Fitzgibbon, A.W.: Automatic video segmentation using spatiotemporal t-junctions. In: BMVC. (2006) 1089–1098