Re-identification of Pedestrians in Crowds Using Dynamic Time Warping

Damien Simonnet, Michal Lewandowski, Sergio A. Velastin,
James Orwell, and Esin Turkbeyler

Digital Imaging Research Centre, Kingston University
Kingston-upon-Thames KT1 2EE, UK
{damien.simonnet,m.lewandowski}@kingston.ac.uk,
sergio.velastin@ieee.org, j.orwell@kingston.ac.uk
Roke Manor Research, Romsey, Hampshire SO51 0ZN, UK
esin.turkbeyler@roke.co.uk

Abstract. This paper presents a new tracking algorithm to solve online the 'Tag and Track' problem in a crowded scene with a network of CCTV Pan, Tilt and Zoom (PTZ) cameras. The dataset is very challenging as the non-overlapping cameras exhibit pan tilt and zoom motions, both smoothly and abruptly. Therefore a tracking-by-detection approach is combined with a re-identification method based on appearance features to solve the re-acquisition problem between non overlapping camera views and crowds occlusions. However, conventional re-identification techniques of multi target trackers, which consist of learning an online appearance model to differentiate the target of interest from other people in the scene, are not suitable for this scenario because the tagged pedestrian moves in an environment where pedestrians walking with them are constantly changing. Therefore, a novel multiple shots re-identification technique is proposed which combines a standard single shot re-identification, based on offline training to recognize humans from different views, with a Dynamic Time Warping (DTW) distance.

1 Introduction

Tracking pedestrians in surveillance videos is an important task, not only in itself but also as a component of pedestrian counting, activity and event recognition, and scene understanding in general. Robust tracking in crowded environments remains a major challenge, mainly due to occlusions and interactions between pedestrians. This paper reports work with a network of Pan, Tilt and Zoom (PTZ) cameras, viewing unstructured crowded scenes [1] where the crowd is relatively free to move over time and space, which corresponds to real scenarios in public spaces such as airports, railway stations or pedestrian streets. A new tracking algorithm is presented to solve on-line the 'Tag and Track' problem, where one pedestrian is nominated (tagged) and then tracked as they are observed from multiple cameras. The dataset[1], illustrated by Fig. 1, is very

[1] Available upon request from the authors.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part I, LNCS 7583, pp. 423–432, 2012.

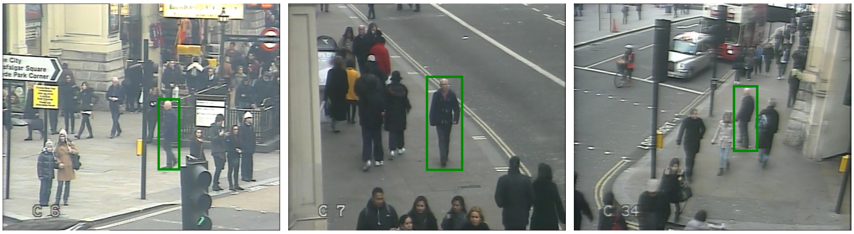


Fig. 1. Network of PTZ cameras dataset. Samples of the person which is tracked with three PTZ cameras (C6, C7 and C34) in a crowded scene during 1 minute and 40 seconds. This dataset is very challenging because the scene is crowded, cameras can move and pedestrians are seen from different non-overlapping viewpoints.

challenging: the non-overlapping cameras exhibit pan tilt and zoom motions, both smoothly and abruptly. Our approach is a novel combination of tracking-by-detection and re-identification using Dynamic Time Warping (DTW) of the resulting appearance features.

The proposed method can be used on arbitrary and diverse input sources: an important advantage is that it does not require the initialization of a background image. Furthermore, it is suitable for integration alongside automatic control of the PTZ cameras, to monitor a scene for the presence of a specific individual. This aspect is not included here. Similarly, a full implementation would require relational details of the cameras in the network to be either specified or learned. These relations will condition the probability density estimate, and are not included here: a uniform prior is assumed.

The rest of the paper is organized as follows. Related work is presented in section 2. Section 3 presents the Tag and Tracker with a network of PTZ cameras. Evaluation is performed in section 4. Finally, the conclusion is given in section 5.

2 Related Work

The proposed approach uses a pipeline process of detection, tracking and re-identification: these aspects are considered in turn below. To detect observations of individuals in a crowded scene, methods based on foreground-background segmentation are not applicable, and are not considered further here. Methods to detect pedestrians in single image frames [2–6] are becoming increasingly accurate, and therefore directly applicable to a (multi-)tracking framework in cluttered conditions. Tracking individuals in a crowded scene must first be distinguished from methods which consider the crowds themselves as global entities [7], which are not considered further here.

Benfold and Reid [8] apply a tracking-by-detection algorithm to data from a high definition fixed camera in a crowded pedestrian street, aiming to obtain a precise location of the pedestrians' heads. First, a combined head/full body detector is applied to detect pedestrians in crowds. This is then combined with a Markov Chain Monte Carlo Data Association (MCMCDA) and a KLT tracker.

The former makes the data association more robust as it is performed over a short time window. The latter uses motion feature, based on appearance, to reinforce the kinematic data association process. For example, the velocity of the target and the observation should be similar, hence many false alarm observations can be removed.

Kuo and Nevatia [9] present a multi-target tracker for crowded scenes from a single camera. First, pedestrian locations in each frame are indicated by a human detector [5]. Then, tracklets (i.e. short sequences of these locations reliably referring to the same individual) are built based on position, size, appearance and motion. These tracklets are then associated using motion, time and an appearance model. This model, learnt with Adaboost, selects the most discriminative pedestrian subparts, using colour (RGB histogram), shape (HOG [2]) and texture (covariance matrix) features. There is a high system latency because the whole sequence is processed before obtaining the results, in contrast to [8].

To re-identify pedestrians, three different types of algorithms can be found in the literature: *short-term*, *contextual long-term*, and *non-contextual long-term* re-identification. *Short-term* re-identification is used by tracking algorithms to associate the data frame by frame and is generally based on appearance features (e.g. colour histogram, texture, edges). *Contextual long-term* re-identification methods [9] use the context of a single static camera to learn online models to differentiate the subject from other pedestrians in the scene and is used to solve the track fragmentation problem. *Non-contextual long-term* re-identification methods [10, 11] are applied across arbitrary cameras and views. The third approach has not yet been applied to the problem of re-identifying pedestrians in a crowded scene with a network of moving cameras. Doretto et al. [10] present such a method for a network of static calibrated cameras in a non crowded environment: this is generalised to non-static, non-calibrated cameras in a crowded environment. In [10], two approaches for re-identification are used, the first called *single shot* bases its re-identification on a single image, and the second called *multiple shots* uses a sequence of images and is therefore more robust.

In this paper, a novel *non-contextual long-term multiple shots* re-identification method is proposed.

3 Tracking in a Network of PTZ Cameras

In this section, a novel algorithm is presented and illustrated by Fig. 2, to solve the 'Tag and Track' requirement. It has three three main steps: human detection, multi-target tracking, and association of the resulting tracklets. First, pedestrians are detected in each frame based on a HOG detector [2]. Then, short reliable tracklets for the pedestrians are built, using spatial multi-target tracker. Then, a *non-contextual long-term* re-identification technique is applied to link tracklets corresponding to the tracked target.

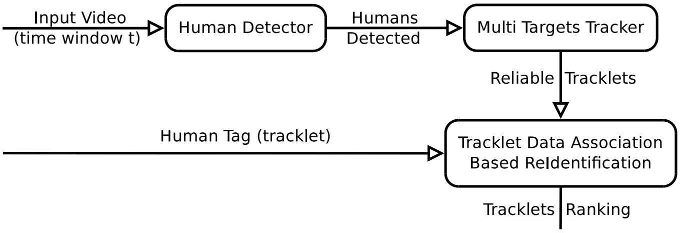


Fig. 2. Flow chart for the Tag and Tracker based on tracklet data association. First, given a tag pedestrian (tracklet), short reliable tracklets are built with a multi-target tracker which uses human detections in each frame during the time window t. Then, a tracklet data association is performed to rank the candidate tracklets. This is intended as one component of a fully automatic system for automatic control of CCTV PTZ cameras.

3.1 Spatial Multi Target Tracker by Detection

As no multi-target tracking methodology prevails for crowded scenes with a moving camera, the Nearest Neighbour Standard Filter (NNSF) [12] is chosen. It is based on a Kalman Filter, consisting of the usual steps: prediction, selection of the validated measurement, and update of the tracker state. Its specificity is the selection of the valid measurement via a Mahalanobis distance.

The measurement and state vector are respectively defined as $Z = \{x, y, h\}$ and $X = \{x, y, h, v_x, v_y, v_h\}$ where x, y represent the position of the pedestrian centre in the image space, h the pedestrian height, and v_x, v_y, v_h are their corresponding speeds. The system transition matrix is defined as a constant velocity model with Gaussian noise to represent acceleration, as this has proved effective for static cameras. As tracking is performed in the image space, the process noise Q and the measurement noise R depend on pedestrian height and are Gaussian.

Then, valid measurements are selected by using the validation volume defined in Eq. (1) where z is a measurement, $S(k+1)$ and $\hat{z}(k+1|k)$ represent respectively the innovation covariance and the measurement prediction, and γ is the normalized innovation square which defines the probability of a measurement to be in the validation volume, based on the Mahalanobis distance. The parameter γ, depending on a probability p to be in the validation volume and the degree of freedom d of the measurement vector, will be used to define what is a reliable tracklet.

$$\mathcal{V}(k+1, \gamma) = \{z,\ [z - \hat{z}(k+1|k)]^t S(k+1)^{-1} [z - \hat{z}(k+1|k)] \leqslant \gamma\} \qquad (1)$$

Finally, the measurement with the smallest Mahalanobis distance is selected to update the state vector. Multiple measurements could have been used for updating the filter. However, the aim of the tracker is to build reliable tracklets and the human detector has removed multiple detections per object during its non maximal suppression stage, so there is no need to use complex methods to update the filter state (e.g. using appearance features to select multiple detected instances of one pedestrian and then updating the state with multiple measurements). It is noted that this selection process can lead to no measurement. In this case the measurement noise is formally infinity, and so the Kalman filter sets the state to be the predicted state and the a posteriori covariance as the a priori covariance.

This NNSF is designed to track one person in clutter, and needs an external initialization to indicate which one person that should be [12]. So, this tracking algorithm needs to be extended for tracking multiple targets whose number is unknown, and to start and stop tracks automatically, and is referred to as the JNNSF (Joint Nearest Neighbour Standard Filter). JNNSF uses a variable number of NNSF filters with a joint data association step, and is composed of three main steps: *track formation*, *track maintenance* and *track termination.*

Track formation occurs when an observation, output by the human detector, has not been associated to any of the current NNSF trackers and when the detection confidence exceeds a high threshold. Then, the *track maintenance* component associates the detection responses with the NNSF trackers. A lower threshold for the detection confidence is used for this stage. The tracker states are first predicted and then updated, using these associated measurements. When no measurement is associated (e.g. due to failure of the detector with occluded pedestrians), the prediction alone is applied. If, after several iterations, the covariance is too big, then a *track termination* step is invoked.

Up to this point, we have presented the underlying detection and tracking process that results in tracklets for different people in the scene. Under favourable circumstances (e.g. low clutter) these tracklets accurately represent the trajectories of each person. However, in many cases, and especially in clutter, these tracklets will be segmented, have swapping ids etc. and a further process is needed as explained in the next section.

3.2 Re-identification of Tracklets across Multiple Shots

Conventional re-identification techniques [8, 9] applied in crowded scenes are not suitable for a multi-camera scenario because they generally learn an on-line model to differentiate the target pedestrian of interest from the other pedestrians in the scene. However, the tagged pedestrian moves in an environment where pedestrians walking with them are also constantly changing. Consequently, a *non-contextual long-term* re-identification method, based on the work of Dikmen et al. [11] (that uses a viewpoint invariant metric based a Large Margin Nearest Neighbour classifier trained on the Viewpoint Invariant Pedestrian Recognition (VIPeR [13]) dataset) is proposed to solve this problem. Moreover, a novel *multiple shots* re-identification technique (more robust than a *single shot* technique which uses only one sample per tracklet) is introduced. Based on the observation that pedestrian walking is a cycling activity which may vary in time and speed, the Dynamic Time Warping (DTW) distance [14] is combined with the similarity distance introduced in [11] to associate tracklets.

More formally, given a target's tracklet T_r, the objective is to identify the most similar subsequent tracklet among a set of N_t candidate tracklets $(T_{c,k})_{k \in [\![1,N_t]\!]}$, generated during a time window t (if $t \to \infty$ then this is a global approach, else a finite t effectively dictates a trade-off between latency and the achievable performance of an on-line system). This is achieved by first extracting and assembling a temporal sequence $F_T = (x_i)_{i \in [\![1, S_T]\!]}$ for each tracklet T where S_T is the size of the tracklet and x_i is an appearance feature obtained from the target at frame i. As a consequence, the similarity of these high dimensional series can be measured by a Dynamic Time Warping distance [14] which minimises the effects of shifting and distortion in time by allowing 'elastic' transformation of series in order to detect similar shapes with different phases.

Given two tracklets (T_a, T_b of respective size S_a and S_b), the re-identification problem is cast as the task of aligning two sequences of observations ($F_a = (x_i)_{i \in [\![1, S_a]\!]}$ and $F_b = (x_j)_{j \in [\![1, S_b]\!]}$) in order to generate the most representative distance measure of their overall difference represented as a a symmetric cost distance matrix $E = \{D(x_i, x_j)\}$ where D is a metric. Afterwards, the algorithm finds the best alignment path (i.e. warping path) between tracklets by satisfying the *boundary*, *monotonicity* and *continuity* conditions which respectively assigns first and last elements of tracklets to each other, preserves the time-ordering of pedestrians, and limits the warping path from long jumps (shifts in time) while aligning tracklets. The final distance between tracklets is given by the end point $P(S_a, S_b)$ of an accumulated global cost matrix P normalised by the sum of tracklets lengths $(S_a + S_b)$, where P has been initialised by Eq. (2) and then updated by a cost function associated based on a warping path defined by Eq.(3).

$$\forall i \in [\![1, S_1]\!], P_{i,1} = \sum_{k=1}^{i} E_{k,1}, \ \forall j \in [\![1, S_2]\!], P_{1,j} = \sum_{k=1}^{j} E_{1,k} \tag{2}$$

$$\forall (i,j) \in [\![2, S_1]\!] \times [\![2, S_2]\!], P(i,j) = \min\{P_{i-1,j-1}, P_{i-1,j}, P_{i,j-1}\} + E_{i,j} \tag{3}$$

In this work, the appearance feature x_i is a vectorized image using 8-bin histograms for each channel of RGB and HSV colour space, but the general approach could be applied to other types of features. To compensate observation differences due to view changes between tracklets, pedestrians are compared with a specialised view-invariant distance metric [11] (distance metric D).

Finally, given any tracklet T_r (corresponding to the tagged target), the multiple shots re-identification is performed by ranking all other tracklets $(T_{c,k})_{k \in [\![1, N_t]\!]}$ within the t seconds time slot according to the distance obtained by the introduced metric. Therefore, the proposed Tag and Tracker is an online tracking algorithm which gives a response with at most t seconds of latency.

4 Evaluation

This section presents an evaluation of the Tag and Track algorithm proposed here on the challenging dataset illustrated in Fig. 1, which involves constantly moving PTZ cameras and contains two complete changes of camera view. The aim of an uncalibrated Tag and Tracker with a network of cameras is not to be able at any time to give the precise location of the pedestrian e.g. in real-world coordinates, but to be able from a starting position (i.e. when a pedestrian is tagged by an operator) to follow the pedestrian through the camera network, localising it in a given camera and to a given image location in that camera.

To our best knowledge, no results have been reported to track a pedestrian in crowded scenes in a network of PTZ cameras. Therefore, here we evaluate the re-identification part our algorithm, because this is key for achieving the aim of the algorithm outlined above. Before going into this aspect, on Fig. 5 we illustrate that the JNNSF tracker is able to build long and reliable tracklets with a moving camera in a crowded scene. Tracklets were built with the normalized innovation square $\gamma = 16.2662$ which corresponds to a probability of true measurement $p = 0.999$ with a measurement degree of freedom $d = 3$ (x, y and h). For the re-identification part, RGB and HSV histograms were extracted from 8×24 rectangular regions, which in turn are densely collected from a regular grid with 4 pixels spacing in vertical and 12 pixels spacing in horizontal direction. This step size is equal to half the width and length of the rectangles resulting in an overlapping representation. The final appearance feature vector x_i is obtained by concatenation of all histograms. Finally, a PCA is applied to obtain a better space for the learning of the metric. The training model for [11] is then retrained with the VIPeR [13] dataset and it uses a vector dimension of 100 after PCA.

The method incorporating Dynamic Time Warping is compared against a single shot re-identification method [11] and its straightforward multiple-shot extension which sums all the similarity scores obtained by [11]. In the rest of the section, we will refer respectively to these three algorithms as *DTW Multiple-Shot*, *Single Shot* and *Simple Multiple-Shot*. The evaluation process is as follows. A pedestrian observation is tagged in a single frame, and the tracker is started with a fixed time window $t = 20s$. The tracklet that includes this observation is designated T_r, and the other tracklets in this window are ranked in order of association. Then, we report the rank of that tracklet which truly represents this pedestrian. However, if the rank is not rank 1, we re-initialize (i.e. re-tag) the correct tracklet, so as not to unduly penalise the method. So ideally, an on-line Tag and Tracker which perfectly satisfies the Tag and Track requirement for a network of cameras will have a rank 1 for each tracklet association.

This evaluation process (tracklet re-identification ranking) for a Tag and Track problem is seen as an extension of Cumulative Matching Characteristics (CMC) [13] used in the evaluation of pedestrian recognition methods with different viewpoints which plots the recognition percentage against the ranking. In the tracklet re-identification evaluation, the height of the bar indicates the ranking of the true tracklet, so a height of 1 indicates a perfect match and higher rankings indicate progressively poorer matching. The results are not displayed as percentages (as in CMC curves) because the number of candidates is not fixed and in fact depends on the number of tracklets generated by the tracker. Consequently, Fig. 3 shows that the DTW Multiple-Shot method significantly outperforms the Single Shot and Simple Multiple-Shot methods and is able to track the target from the first frame tagged until the end of the

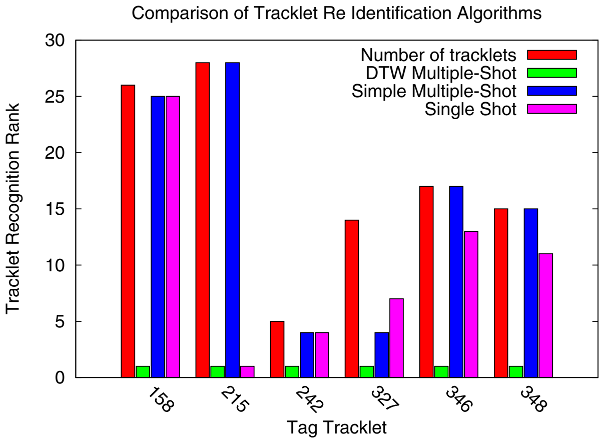
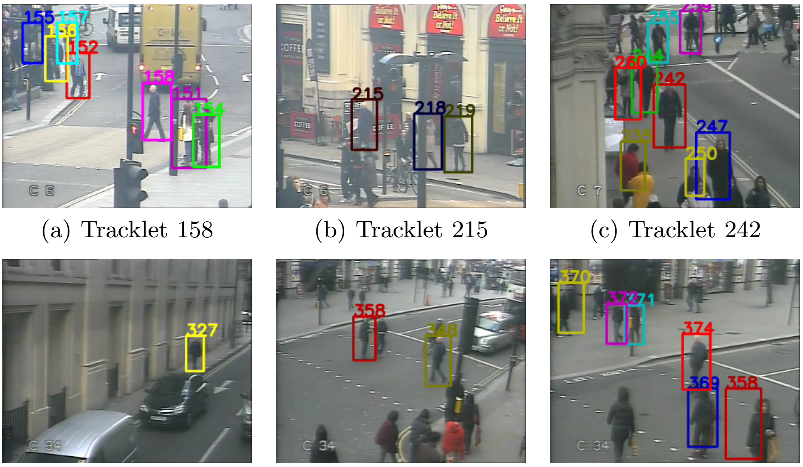


Fig. 3. Comparison of the three tracklet re-identification methods: DTW Multiple-Shot (ours), Single Shot and Simple Multiple-Shot. The JNNSF multiple target tracker was able to build automatically seven independent tracklets (158, 215, 242, 327, 346, 348 and 374) corresponding to the tracked target (plus many others corresponding to other people). The number of tracklets corresponds to the total number of candidate tracklets generated by the JNNSF tracker in a time window of $t = 20s$ and it also represents the worse possible ranking, the best being 1. Our method outperformed the others and is able to track the tagged target from the beginning to the end of the scene.



(a) Tracklet 158 (b) Tracklet 215 (c) Tracklet 242

(d) Tracklet 327 (e) Tracklet 348 (f) Tracklet 374

Fig. 4. An illustrative time sequence to show that the Tag and Track algorithm is able to link tracklets in a challenging scenario: it first links tracklets 158 and 215, (a) and (b) from the same camera after significant occlusion and challenging camera move. Then it also links tracklets between different camera (i.e. tracklet 215, 242 and 327, (b)-(d)) and also links tracklets after occlusion when the camera is moving (i.e. tracklet 348 and 374, (e) and (f)).

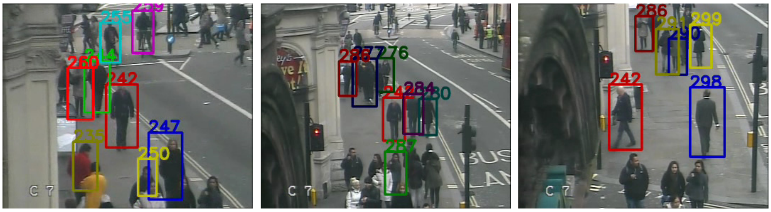


Fig. 5. In this illustrative example, tracklet 242 is reliably built within the crowds from the entrance of the target to the field of view (FOV) of this camera (C7) until the target leaves the FOV. Note how the position and zoom of the camera is changing all the time.

sequence. Although the results against the Single Shot method were expected, the comparison against the Simple Multiple-Shot method shows that extending a Single Shot method is not straightforward in a challenging dataset with crowds and moving cameras, and in fact Simple Multiple Shots provides worse results than the Single Shot.

As it has been seen for the Single Shot and even for the Simple Multiple-Shot approach, a comparison which is based purely on appearance features is not discriminative enough to perform robust tracklet re-identification. This is due not only to very low variability in the colour space of the different people but particularly to lack of structural information within tracklets. To overcome that, the process of appearance based re-identification has been extended by taking into account the temporal structure of tracklets. As a result, the proposed DTW Multiple-Shot method proves to be significantly more effective than the other two basic approaches, achieving first-rank classification in all cases.

An example of tracklets association is given in Fig. 4 to illustrate the ability of the approach to deal with challenging situations.

5 Conclusion

This paper has presented a novel tracker to address the Tag and Track problem in a network of PTZ cameras based on three main components: a backgroundless people detector, a multi-target tracker and a multiple shots re-identification technique. In addition, a new methodology to compare Tag and Track tracking in a network of cameras is introduced. Finally, experiments are conducted on a new type of dataset (non-overlapping PTZ camera network with zoom motions, both smoothly and abruptly) where the algorithm presented shows very good results mainly due to the introduction of the Dynamic Time Warping distance in the re-identification method.

As future work, results can be improved along three different axes: human detection in crowds (usage of part based detectors), re-identification (combine DTW distance with [15] which recently shows better results than [11]) and use of motion in a fully automatic system for a network of PTZ cameras.

Acknowledgements. The first author is grateful to Roke Manor Research for their financial and technical support. The authors are grateful to the Seventh Framework Programme 'ADDPRIV' (No. 261653) which partly funded this work.

References

1. Rodriguez, M., Ali, S., Kanade, T.: Tracking in unstructured crowded scenes. In: 12th International Conference on Computer Vision, pp. 1389–1396. IEEE (2009)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, vol. 1, p. 886 (2005)
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 75(2), 247–266 (2007)
4. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1627–1645 (2009)
5. Huang, C., Nevatia, R.: High performance object detection by collaborative learning of joint ranking of granule features. In: CVPR, pp. 41–48 (2010)
6. Duan, G., Ai, H., Lao, S.: A Structural Filter Approach to Human Detection. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 238–251. Springer, Heidelberg (2010)
7. Ali, S., Shah, M.: Floor Fields for Tracking in High Density Crowd Scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 1–14. Springer, Heidelberg (2008)
8. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: Computer Vision and Pattern Recognition, pp. 3457–3464 (2011)
9. Kuo, C., Nevatia, R.: How does person identity recognition help multi-person tracking? In: CVPR, pp. 1217–1224. IEEE (2011)
10. Doretto, G., Sebastian, T., Tu, P., Rittscher, J.: Appearance-based person reidentification in camera networks. Journal of Ambient Intelligence and Humanized Computing 2, 127–151 (2010)
11. Dikmen, M., Akbas, E., Huang, T.S., Ahuja, N.: Pedestrian Recognition with a Learned Metric. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part IV. LNCS, vol. 6495, pp. 501–512. Springer, Heidelberg (2011)
12. Bar-Shalom, Y., Li, X.: Multitarget-multisensor tracking: principles and techniques. Yaakov Bar-Shalom (1995)
13. Gray, D., Tao, H.: Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008)
14. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall, Inc. (1993)
15. Tatsuo Kozakaya, S.I., Kubota, S.: Random ensemble metrics for object recognition. In: IEEE International Conference on Computer Vision (2011)