2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance

Online Multiperson Tracking With Occlusion Reasoning and Unsupervised Track Motion Model

Niall McLaughlin, Jesus Martinez Del Rincon and Paul Miller
Centre for Secure Information Technologies (CSIT)
Queen's University Belfast
{n.mclaughlin,j.martinez-del-rincon,p.miller}@qub.ac.uk

Abstract

We address the problem of multi-target tracking in realistic crowded conditions by introducing a novel dual-stage online tracking algorithm. The problem of data-association between tracks and detections, based on appearance, is often complicated by partial occlusion. In the first stage, we address the issue of occlusion with a novel method of robust data-association, that can be used to compute the appearance similarity between tracks and detections without the need for explicit knowledge of the occluded regions. In the second stage, broken tracks are linked based on motion and appearance, using an online-learned linking model. The online-learned motion-model for track linking uses the confident tracks from the first stage tracker as training examples. The new approach has been tested on the town centre dataset and has performance comparable with the present state-of-the-art.

1. Introduction

Multi-target tracking is an important component of various applications in computer vision such as visual surveillance and sports analysis. However, despite its crucial role, consistently and accurately tracking multiple people over time remains a challenge. This is due to the many sources of uncertainty, e.g., measurement noise, background clutter, changing background and illumination conditions, significant occlusions and distractors. Additional difficulties arise from the particularities of dealing with multiple targets, such as appearance similarity between tracked objects, overcrowded scenes, monocular camera and complex interactions between targets.

This paper addresses the problem of detection and tracking multiple people in cluttered and overcrowded scenes using a monocular camera. It relies on a robust part-based human detector and a dual-stage tracking-by-detection framework to solve the inherent ambiguity of the tracking problem. This approach is a response to the fact that targets can be occluded for long periods of time, making both detection and association challenging.

To solve this problem our framework employs a two-stage strategy. In the first stage, occlusions are taken into account when solving the data-association problem for linking tracks with detections. In the second stage, broken tracks are linked using a motion-model learned online, taking into account temporal and geometric constraints.

The rest of the paper is organised as follows: In Section 2 we describe the first stage of our proposed online tracking-by-detection algorithm. Section 3 describes the novel approach for dealing with occlusion. In Section 4 we explain our second tracking stage that uses an online-learned motion-model for track linking. In Section 5 we discuss the experimental results and in Section 6 we present our conclusions.

1.1. State of the Art

Initial attempts to deal with multi-target tracking were based on the Markovian assumption. Sequential Monte Carlo methods [11, 14] provide a theoretical framework to model and integrate multiple sources of uncertainty, considering only the information from past frames. In practice these methods are limited to a few simultaneous targets due to the curse of dimensionality or the difficulty of designing appropriate interaction models.

Recently, significant advances in pedestrian detectors [6, 7] have made tracking-by-detection approaches practical [9, 20]. Under real-world circumstances these methods can suffer from the temporal delay introduced by the detector itself, as well as the need to optimise the tracking solution over a long temporal sequence [2, 10].

Few tracking-by-detection approaches have targeted online tracking or short temporal sliding windows [4, 1]. The common problem faced by these approaches is the limited reasoning capabilities of the optimisation stage due to the short time-span of the window. Proposed solutions have

978-1-4799-0703-8/13/$31.00

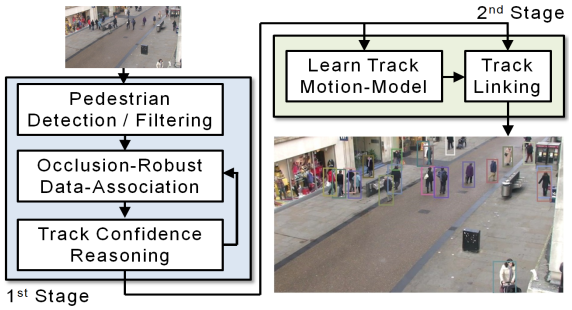

Figure 1. Flowchart of our tracking system.

included particle-filtering [4] or a combination of tracking-by-detection with other trackers such as KLT [1]. The goal of these approaches is to generate short confident tracks known as tracklets [1]. However, this approach relies on good detector performance [4], which may not be the case in more complex sequences. This shortcoming can be addressed by incorporating temporal context to reduce the number of false-positives and missing detections [5].

In this paper, we propose a tracking-by-detection framework which fits with the last category. The framework is able to balance both tracking and data association stages in a suitable manner. Our novel contributions include the use of a posterior union model (PUM) to discount occluded features during data association, and the use of an online learned track motion-model to calculate cost functions for efficient tracklet re-association.

2. Tracking-by-detection framework

A block diagram of our tracking approach is depicted in Fig.1. There are five main parts: detection and filtering, occlusion-robust data-association, track confidence reasoning, learning of the motion model, and track linking. First a pedestrian detector is applied. The detections are filtered and then associated with the corresponding tracks, taking into account occlusion. Track confidence is assessed and the confident tracks are passed to the second stage, where a track motion model is learned and used to link broken tracks.

2.1. Pedestrian Detection

In each frame, pedestrians are detected using the Poselets pedestrian detector [3] . The Poselets detector is a parts-based pedestrian detector, where the extraction of the pedestrian bounding boxes is not the primary goal but a consequence of the sum of detections of specific configurations of the body parts, making Poselets robust against extensive and long term occlusion. Hence, this choice of detector is coherent with the expected uncertainty of overcrowded scenes, where occlusions are not the exception but the rule. Additionally, it was shown in [3] that the Poselets pedestrian detector has higher accuracy than the popular parts-based detector of [7].

Although it is possible to adjust the parameters of the detector to reduce the number of false-positives, we choose to retain all candidate detections following similar reasoning to [4]. Thus, we retain all true-positive detections, even those with low confidence scores, at the expense of a large number of false-positive detections that must be filtered by other means.

To reduce the number of false-positives we use camera-calibration and the known distribution of human heights [21], removing detections outside the range $1.4 - 2$m, based on the assumption that all detections are of upright pedestrians standing on the same ground-plane. Each detection is projected into world-space and its true-height estimated from its world-position and bounding box height [15]. This step removes the vast majority of false-positive detections. Remaining false-positives are dealt with during the track initialisation stage (see Section 2.2) based on the differing motion characteristics of true-positive and false-positive tracks.

In addition, we extend the Poselets output to segment the main body areas from the background by considering the corresponding Poselet activations. This allows us to model the appearance of each person, using the 3D colour histograms of the main body areas (see Section 3).

2.2. Track Generation

The track generation mechanism is an online tracker based on the Markovian assumption that the state at time t depends only on the state at time $t-1$ and the observation at time t. A Kalman filter is used to better estimate the state of each track at each time-step, given the set of noisy detections previously associated with the track. The state vector of each track in the Kalman filter consists of its world-position, velocity, bounding-box size and rate-of-change of bounding-box size. At each time-step the Kalman filter predicts a smoothed estimate of the track's state using the observation at time t and the state at time $t-1$.

2.2.1 Data Association

At each time-step the online tracker receives a new set of detections, each of which is either associated with an existing track or used to initialise a new track.

Similarly to the approach used in [20], the Markovian tracker uses a track hierarchy consisting of confident and non-confident tracks. Confident tracks have been associated with many detections and passed several initialisation requirements, while non-confident tracks have been associated with a smaller number of detections and have not passed the initialisation requirements to be considered confident tracks.

Data association between new detections and tracks is performed hierarchically. First, association is attempted between new detections and confident tracks. Association is then attempted between any remaining unassociated detections and non-confident tracks. Finally, any still remaining detections are used to initialise new non-confident tracks. In effect, this hierarchical data association scheme weights association decisions in favour of tracks with more supporting evidence.

The cost of linking track $\mathbf{t}_i$ and detection $\mathbf{d}_j$ is defined as

$$C(\mathbf{t}_i, \mathbf{d}_j) = L(\mathbf{t}_i, \mathbf{d}_j)B(\mathbf{t}_i, \mathbf{d}_j)D(\mathbf{t}_i, \mathbf{d}_j) \quad (1)$$

where $L(\mathbf{t}_i, \mathbf{d}_j)$ is the appearance similarity between track $\mathbf{t}_i$ and detection $\mathbf{d}_j$, $B(\mathbf{t}_i, \mathbf{d}_j)$ is the size difference between the bounding boxes, and $D(\mathbf{t}_i, \mathbf{d}_j)$ is the euclidean distance between the predicted centre position of the track $\mathbf{t}_i$ and the centre position of detection $\mathbf{d}_j$. An appearance model is maintained for each track; calculation of the appearance similarity $L(\mathbf{t}_i, \mathbf{d}_j)$, which takes into account the possibility of partial occlusion, is discussed in Section 3. The size difference between the bounding boxes is calculated as $(1 - \left\| \frac{b_d}{b_t} \right\|)^2$ where b_t is the length of the diagonal of the bounding box of track $\mathbf{t}_i$ and b_d is the length of the corresponding diagonal of detection $\mathbf{d}_j$.

Association between tracks and detections is modelled as a linear assignment problem (LAP) where each detection may only be associated with a single track. The optimal minimum cost solution to this assignment problem can be efficiently computed using the Hungarian algorithm [12]. An assignment matrix between tracks and detections is created, where the cost of assigning detection d_i to track t_j is

$$A(\mathbf{d}_i, \mathbf{t}_j) = \begin{cases} C(\mathbf{t}_i, \mathbf{d}_j) & D(\mathbf{t}_i, \mathbf{d}_j) < \lambda \\ \infty & D(\mathbf{t}_i, \mathbf{d}_j) \geq \lambda \end{cases} \quad (2)$$

where $C(\mathbf{t}_i, \mathbf{d}_j)$ is the cost of linking detection d_i and track t_j as defined in Eq. (1), and where $D(\mathbf{t}_i, \mathbf{d}_j)$ is the euclidean distance between the predicted centre location of track $\mathbf{t}_i$ and the centre location of detection $\mathbf{d}_j$. By limiting associations to within a radius λ of the predicted location of each track, false associations due to detector failure and coincidental appearance similarity are reduced.

2.2.2 Track Confidence Reasoning: Promotion and Termination

All tracks are initialised as non-confident tracks. After passing a series of tests, they may be promoted to confident tracks. The tests performed are as follows: firstly, the number of detections associated with the non-confident track must be greater than a constant α. Secondly, the average number of detections associated with the track during a time-window, of duration β frames, must be greater than γ. Finally, the speed of the track during the window β must be greater than 0. The parameters α and β are optimised using 10-fold cross-validation (see Section 5).

To cope with short-term occlusions and temporary detector failure, confident tracks are allowed to 'drift' for a short time-period without being associated to a detection. During this period the track position is predicted using the Kalman filter. Tracks are terminated if they have not been associated with a detection for a long time-period, causing track confidence to drop, or if they have reached the edge of the field of view.

3. Occlusion-Robust Appearance Similarity

Appearance is one of the most discriminative features that can be used to resolve tracking ambiguity when spatio-temporal features are not sufficient, such as can happen in overcrowded scenes. However, data-association between tracks and detections based on appearance can be difficult due to the interaction of people with background objects and other pedestrians, which can lead to occlusion.

To increase tracking accuracy, we use the posterior union model (PUM) [17] to calculate the appearance similarity between tracks and detections while taking the possibility of partial occlusion into account. This missing feature method has previously been used for speaker identification in noisy conditions [17], and facial recognition given partial occlusion [16]. Missing feature methods are used to calculate a matching score, between a model and a partially corrupted object, by ignoring the contribution of the corrupted parts to the overall score. The novelty of the PUM is that such a score can be calculated without the need for explicit knowledge of the corrupted parts. This makes it useful for tracking applications where it can be difficult to accurately identify the occluded regions.

Assume that both the detection $\mathbf{d}_j$ and track $\mathbf{t}_i$ can each be represented by sets of corresponding parts. In a conventional approach the matching scores of all corresponding parts would be combined to produce an overall similarity score between $\mathbf{d}_j$ and $\mathbf{t}_i$. However, given partial occlusion, some of the matching scores will be corrupted. The PUM finds the optimal subset of matching scores, i.e. the subset with maximum posterior probability, thus computing an overall appearance similarity while effectively ignoring the occluded parts.

Let $\mathbf{t}_i = (t_i^1, t_i^2, \ldots, t_i^n)$ represent the appearance model for track $\mathbf{t}_i$, composed of n colour histogram features extracted from a grid of non-overlapping blocks covering the main body areas, and let $\mathbf{d}_j = (d_j^1, d_j^2, \ldots, d_j^n)$ represent the corresponding appearance model of detection $\mathbf{d}_j$. Given a subset of the features $X_s \subseteq [1 \ldots n]$, we define the appearance similarity $Q(\mathbf{d}_j|\mathbf{t}_i, X_s)$ between $\mathbf{d}_j$ and $\mathbf{t}_i$ as

$$Q(\mathbf{d}_j|\mathbf{t}_i, X_s) = \prod_{s \in X_s} M^{b(t_i^s, d_j^s)} \quad (3)$$

where M is a positive base number and $b(t_i^s, d_j^s)$ is the Bhattacharyya distance between the appearance features t_i^s and d_j^s. In a realistic tracking scenario, some of the appearance features may be corrupted by occlusion, meaning the optimal feature subset X_s' is unknown. We therefore define the overall similarity between detection $\mathbf{d}_j$ and track $\mathbf{t}_i$ as

$$L(\mathbf{d}_j, \mathbf{t}_i) \propto \max_{X_s'} \frac{Q(\mathbf{d}_j|\mathbf{t}_i, X_s')}{\sum_{\mathbf{t}'} Q(\mathbf{d}_j|\mathbf{t}', X_s')} \quad (4)$$

where $\mathbf{t}'$ is the set of all tracks including $\mathbf{t}_i$. This posterior-like function computes the overall similarity of detection $\mathbf{d}_j$ with track $\mathbf{t}_i$ by maximizing over all possible feature subsets X_s' to find the optimal feature subset containing the reliable unoccluded features. A fast approximate algorithm for computing the optimal feature subset is discussed in [17].

We represent the appearance model of each track using 3D colour histograms extracted from a grid of non-overlapping blocks covering the head, torso and upper legs. The lower legs were not included in the appearance model due to their small size and motion, which resulted in background information being included in the model.

To cope with appearance variation, each track's appearance model is updated using η models retained from previously associated detections. The appearance model is calculated as the median of all the retained features, allowing the model to cope with short-term occlusions and temporarily incorrect associations, while remaining adaptive to long-term appearance variation.

4. Unsupervised Track Linking

A problem occurs for online multi-target trackers when a track is incorrectly broken, as can happen when a track experiences long-term occlusion, or when the pedestrian detector fails for an extended number of frames. In such cases, while it is possible that the person may be reacquired, they will be incorrectly labeled and their location during the period between the two tracks will be lost. In the second stage of our tracker, we address this problem by introducing an additional track-linking stage that uses an online-learned track motion-model to link such broken tracks.

Many association algorithms have been proposed in the literature such as, Multiple Hypothesis Tracking[19], Joint Probabilistic Data Association Filter[8], or approaches based on the Hungarian algorithm [12]. However, these approaches typically suffer from an explosion in the size of the hypothesis space given a large number of targets. Alternatively, as demonstrated by [5, 22], if the interaction and track-linking cost functions are properly designed, the association problem can be solved with only a simple schema.

We therefore focus on taking advantage of the output from the first tracking stage for filtering missed detections and false-positives as well as providing useful information for re-associating broken tracks. This additional tracking stage works in conjunction with the first tracking stage to re-associate broken tracks. Unlike most tracklet association approaches, the re-associaiton model is learned in an unsupervised manner, using the output of the first stage online Markovian tracker as training data.

4.1. Track Re-association Cost Function

After the tracker initialisation period, we assume that persons may only enter or exit the scene from specific predefined regions, such as near the image edge. Therefore, when a track is terminated in a region where the cause is not likely to be the person's physical exit from the scene, we retain a record of the track for a time window of up to ζ frames. When a non-confident track later passes the requirements to be promoted to the status of confident track (see Section 2.2.2), a cost function is evaluated to determine if it is likely to be a continuation of a previously terminated track. The cost of linking newly initialised confident track $\mathbf{t}_n$ with previously terminated track $\mathbf{t}_k$ is defined as

$$C(\mathbf{t}_n, \mathbf{t}_k) = M(\mathbf{t}_n, \mathbf{t}_k, \phi_1)K(\mathbf{t}_n, \mathbf{t}_k, \phi_2)L(\mathbf{t}_n, \mathbf{t}_k)T(\mathbf{t}_n, \mathbf{t}_k) \quad (5)$$

where $M(\mathbf{t}_n, \mathbf{t}_k, \phi_1)$ is the cost of the motion discrepancy between the tracks, $K(\mathbf{t}_n, \mathbf{t}_k, \phi_2)$ is the cost of the angular discrepancy between the tracks, $L(\mathbf{t}_n, \mathbf{t}_k)$ is the appearance similarity as defined by Eq.4, and $T(\mathbf{t}_n, \mathbf{t}_k)$ is a time penalty defined as $1 - \frac{\Delta T}{\zeta}$, where ΔT is the time difference between the end of track $\mathbf{t}_k$ and the start of track $\mathbf{t}_n$. To compute the motion and angular discrepancy costs between tracks $\mathbf{t}_n$ and $\mathbf{t}_k$, the trajectory of $\mathbf{t}_k$ is predicted forwards for ΔT s assuming linear motion. The motion discrepancy cost $M(\mathbf{t}_n, \mathbf{t}_k, \phi_1)$ is based on $\|\mathbf{t}_n - \mathbf{t}_k\|_2$ the euclidean distance between the predicted position of $\mathbf{t}_k$ at time $t+\Delta$T s and the starting position of $\mathbf{t}_n$. The distribution of motion discrepancy distances between the predicted position of track $\mathbf{t}_k$ and the starting position of track $\mathbf{t}_n$ is modelled as a zero-mean Gaussian defined as $M = e^{-\left(\frac{\|\mathbf{t}_n - \mathbf{t}_k\|_2^2}{\phi_1^2}\right)}$, where the parameter ϕ_1 represents the variance of the distribution. The angular discrepancy cost $K(\mathbf{t}_n, \mathbf{t}_k, \phi_2)$ is based on $\Theta_{k,n}$ the angle between the predicted trajectory of $\mathbf{t}_k$ and the smoothed trajectory of $\mathbf{t}_n$. The distribution of angular discrepancies between the trajectory of $\mathbf{t}_k$ and the trajectory of $\mathbf{t}_n$ is modelled as a zero-mean Gaussian, defined as $K = e^{-\left(\frac{\Theta_{k,n}{}^2}{\phi_2^2}\right)}$, where the parameter ϕ_2 represents the variance of the distribution.

Based on the above cost function, the association between new tracks and previous terminated tracks is optimised using the Hungarian algorithm over a short sliding window.

4.2. Automatic Parameter Learning

The parameters ϕ_1 and ϕ_2 of the track linking cost functions are learned online using statistics collected from the confident tracks produced by the main tracker. Only the set of confident tracks are used for parameter learning as these tracks are more likely to reflect the true distribution of target motion than the non-confident tracks. In effect, by using the output of the Markvoian tracker to learn the parameters of a track linking model, the system is generating its own training data.

The parameters ϕ_1 and ϕ_2 are learned by breaking the confident tracks at random locations. Let $\mathbf{t}_b$ be the confident track under consideration. In order to update the parameter ϕ_1, used in modelling the motion discrepancy between tracks, a point p along $\mathbf{t}_b$ is randomly selected. A Kalman filter is then used to linearly project the path of $\mathbf{t}_b$ forward in time from p for a random duration $\Delta T'$ where $\Delta T' \leq \zeta$. The Euclidean distance between the predicted position of $\mathbf{t}_b$ after $\Delta T'$ s, assuming linear motion, and the actual position of $\mathbf{t}_b$ after $\Delta T'$ s is recorded. The distribution of displacements is modelled as a zero-mean Gaussian, where the measured distances between the predicted and actual track positions are used to update the variance parameter ϕ_1.

The parameter ϕ_2, which models the angular discrepancy between tracks, is updated in a similar manner. A Kalman filter is used to compute the instantaneous trajectory of $\mathbf{t}_b$ at points p and $p + \Delta T'$, assuming linear motion. The angle between the trajectory of $\mathbf{t}_b$ at p and its trajectory at $p+\Delta T'$ is recorded. The distribution of angles is modelled as a zero mean Gaussian, and the recorded angle between the track trajectory at both time instants is used to update variance parameter ϕ_2.

5. Experimental Evaluation

The tracker was evaluated on the town centre dataset [1], which contains a realistic street scenario captured with a resolution of 1920×1080 pixels at 25 fps. It features naturalistic pedestrian behaviour, with many cases of short-term partial and full occlusions, as well as several cases of long-term occlusion. The crowd density varies from sparse to moderately crowded. The total number of frames with ground-truth in this dataset is 4500. This sequence has been used extensively in the recent tracking papers, facilitating comparison with the state of the art.

The track-initialisation parameters α and β, which specify the number of detections and the detection-rate a non-confident track must achieve before being considered confident, were set using 10-fold cross-validation. The ground-truth information was split into 10 non-overlapping parts and the parameters were optimised on each part individually. The tracker was then tested on the whole sequence using the optimised parameters. We report the mean and standard deviation of the results from these runs. The other parameters: λ, the search radius for track-detection association, γ, the window length for track-initialisation, ζ, the window-length for track linking, and η, the number of appearance models retained by each track, were set to: 50 pixels, 25 frames, 75 frames and 25 models, respectively.

5.1. Comparison with the Literature

Shown in Table 5.1 are comparisons of our system with state-of-the-art trackers. We use the standard CLEAR MOT performance metrics, and the PASCAL 50% overlap criterion for associating tracks with the ground-truth [1]. Results are shown for the system tested on all 4500 frames of the town centre dataset.

From Table 5.1 we can observe that the TA results from our approach are better than, or comparable with the literature, even against techniques based on global optimisation [24, 9], or including more complex reasoning, such as social behaviour [13, 18, 23]. TA is widely accepted as a good reflection of true tracker performance, as it measures false-positives, false-negatives and ID-switches, whereas TP simply measures how closely the tracker follows the ground-truth regardless of any other errors.

5.2. Partial Occlusion

We test our novel method of computing occlusion robust matching scores between tracks and detections by varying the appearance models used to represent each track / detection and by varying appearance similarity method used during the data-association step.

In the first part of this experiment each track's appearance model was represented using the 3D colour histogram of the whole bounding box, replicating a simple tracker. The appearance similarity between tracks and detections was calculated using the Bhattacharyya distance.

Next a parts-based appearance model was used. This appearance model comprised the 3D colour histograms of non-overlapping blocks extracted from the head, torso and upper legs. The head was represented by a single histogram due to its small size, while the torso and upper legs were each represented by the colour histograms from 4 non-overlapping blocks. The appearance similarity between a given track-detection pair was first calculated as the product of all the individual similarity scores from the corresponding parts, replicating a simple parts-based approach. Then the appearance similarity was calculated using the PUM to select the optimal subset of non-occluded features. The results of this experiment are shown in Table 5.2.

From the results in Table 5.2 it can be concluded that not only does using a parts-based appearance model improve the tracker performance, but our novel mechanism for automatically handling partial occlusion, gives an additional improvement in performance.

Method	TA	TP	Precision	Recall
Our method	74.15±0.69	72.41±0.03	90.4±1.29	83.27± 0.76
H. Izadinia *et al*. [9]	75.7	71.6	93.6	81.8
B. Benfold *et al*. [1]	61.3	80.3	82.0	79.0
G. Shu *et al*. [20]	72.9	71.3	-	-
K. Yamaguchi *et al*. [23]	61.3	70.9	71.1	64.0
S. Pellegrini *et al*. [18]	63.4	70.7	70.8	64.1
L. Zhang *et al*. [24]	65.7	71.5	71.5	66.1
L. Leal-Taixe *et al*. [13]	67.3	71.5	71.6	67.6

Table 1. Comparison of the results produced by our system with the literature.

Method	TA	TP	Pre.	Rec.
Posterior Union Method	74.15	72.41	90.40	83.27
Product of All Scores	72.61	72.31	87.58	84.96
Whole Bounding Box	70.54	72.21	86.68	83.59

Table 2. Results using occlusion robust appearance similarity.

6. Conclusions

In this paper we have introduced a novel online dual-stage multi-target tracking framework, that is capable of handling partial occlusions and of linking broken tracks. The system includes a novel occlusion-robust method for calculating the appearance similarity between tracks and detections, that does not require explicit identification of the occluded regions. We have also demonstrated a novel method for online learning of a track linking motion model. The performance of this system has been shown to be better than, or comparable with, the present state-of-the-art.

References

[1] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, 2011.

[2] J. Berclaz, F. Fleuret, and P. Fua. Multiple object tracking using flow linear programming. In *PETS-Winter*, 2009.

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[4] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Robust tracking-by-detection using a detector confidence particle filter. In *ICCV*, 2009.

[5] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE Trans. PAMI*, 33(9):1820–1833, 2011.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. PAMI*, 32(9):1627–1645, 2009.

[8] T. Fortmann, Y. Bar-Shalom, and M. Scheffe. Multi-target tracking using joint probabilistic data association. In *IEEE CDC*, 1980.

[9] H. Izadinia, I. Saleemi, W. Li, and M. Shah. (mp)2t: multiple people multiple parts tracker. In *ECCV*, 2012.

[10] X. Jiang, E. Rodner, and J. Denzler. Multi-person tracking-by-detection based on calibrated multi-camera systems. In *ICCVG*, 2012.

[11] Z. Khan, T. Balch, and F. Dellaert. An mcmc-based particle filter for tracking multiple interacting targets. In *ECCV*, 2004.

[12] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

[13] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *ICCV*, 2011.

[14] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. *IJCV*, 19(1):57–71, 2000.

[15] J. Martínez-del Rincón, J. E. Herrero-Jaraba, J. R. Gómez, and C. Orrite-Urunuela. Automatic left luggage detection and tracking using multi-camera ukf. In *PETS*, 2006.

[16] N. McLaughlin, J. Ming, and D. Crookes. Robust bimodal person identification using face and speech with limited training data and corruption of both modalities. In *Interspeech*, 2011.

[17] J. Ming, T. Hazen, J. Glass, and D. Reynolds. Robust speaker recognition in noisy conditions. *IEEE Trans. ASLP*, 15(5):1711–1723, 2007.

[18] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *IEEE CV*, 2009.

[19] D. B. Reid. An algorithm for tracking multiple targets. *IEEE Trans. on Auto. Con.*, 24:843–854, 1979.

[20] G. Shu, A. Dehghan, and O. Oreifej. Part-based multiple-person tracking with partial occlusion handling. In *CVPR*, 2012.

[21] P. M. Visscher. Sizing up human height variation. *Nature genetics*, 40(5):489–490, 2008.

[22] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, 2005.

[23] K. Yamaguchi, A. Berg, L. Ortiz, and T. Berg. Who are you with and where are you going? In *CVPR*, 2012.

[24] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008.