To Track or To Detect? An Ensemble Framework for Optimal Selection

Xu Yan, Xuqing Wu, Ioannis A. Kakadiaris, and Shishir K. Shah*

Department of Computer Science, University of Houston
Houston, TX 77204-3010, USA
xyan5@uh.edu, xuqingwu9@gmail.com, ioannisk@uh.edu, sshah@central.uh.edu

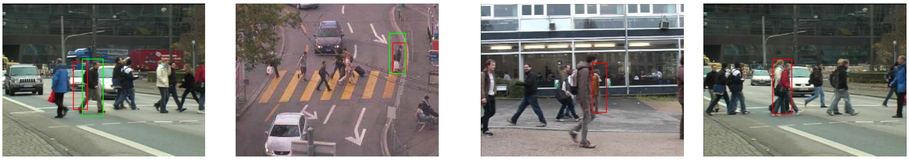
Abstract. This paper presents a novel approach for multi-target tracking using an ensemble framework that optimally chooses target tracking results from that of independent trackers and a detector at each time step. The ensemble model is designed to select the best candidate scored by a function integrating detection confidence, appearance affinity, and smoothness constraints imposed using geometry and motion information. Parameters of our association score function are discriminatively trained with a max-margin framework. Optimal selection is achieved through a hierarchical data association step that progressively associates candidates to targets. By introducing a second target classifier and using the ranking score from the pre-trained classifier as the detection confidence measure, we add additional robustness against unreliable detections. The proposed algorithm robustly tracks a large number of moving objects in complex scenes with occlusions. We evaluate our approach on a variety of public datasets and show promising improvements over state-of-the-art methods.

1 Introduction

Visual tracking of multiple targets in complex scenes captured by a monocular, potentially moving, and uncalibrated camera is a very challenging problem due to measurement noise, cluttered-background, uncertainty of the target motion, occlusions, and illumination changes [1]. While traditional methods for tracking have focused on improving the robustness of motion models and predictive filters, recent advances in methods for object detection [2–4] have led to the development of a number of *tracking-by-detection* [5–12] approaches. These methods first apply a learned discriminative model to detect objects in each frame independently, and then associate detections across frames to identify each object's unique spatio-temporal trajectory. However, varying visual properties of the object of interest often results in false positives and missed detections. Hence, the

* This work was supported in part by the US Department of Justice, grant number 2009-MU-MU-K004. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of our sponsors.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part V, LNCS 7576, pp. 594–607, 2012.



(a) Tracking drift (b) Imprecise detection (c) Missing detection (d) Occlusion

Fig. 1. Examples of output from a tracker (in red box) and a detector (in green box)

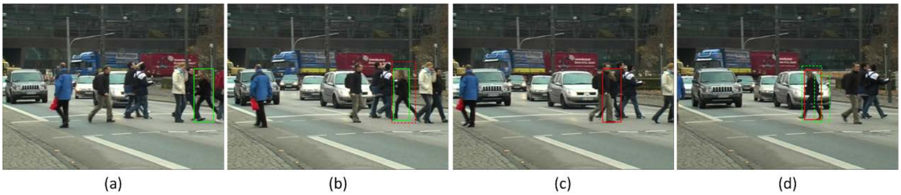


Fig. 2. Using output from both the tracker and the detector, our algorithm selects the best candidate and associates it to the tracked target. Candidates from the tracker are shown in red boxes and candidates from the detector are in green boxes. Solid boxes represent the best candidate selected by the algorithm.

resulting association problem has to be resolved by inferring between-object interactions using incomplete data sets. Several approaches have been proposed to address this problem by optimizing detection assignments in the spatio-temporal context [13, 14, 11, 15], while other methods have focused on achieving the necessary precision by coupling a robust tracker that can update its predictive model (motion and object attributes) guided by the detection confidence and discriminative features obtained from multiple cues [9, 7, 16]. The improved tracking performance reported by these methods indicates that such a combination is desirable. Nonetheless, the positive or negative contribution of the chosen predictive model and the detector at each time step within the combination term is not well understood. Further, it is not guaranteed that each term will have an equivalent contribution towards tracking a target and the weighting parameters chosen empirically could deteriorate the tracking result in previously unobserved scenarios.

Traditional trackers that depend on the appearance model and motion prediction perform poorly in the presence of abrupt motion changes and cause template drifts. The true target gradually shifts away from the tracking template because of the error in the motion model. In addition, tracking drifts and photometric variations make it hard to maintain unique identities among targets and cause frequent identity switches. On the other hand, detection results suffer from long-term occlusion, dynamic backgrounds and low-resolution images. Since output from either the detector or the tracker can be sparse and unreliable, one solution to alleviate the problem is to create an abundant number of potential candidates to increase the probability of finding a more accurate candidate for the target of interest. For example, the visual tracker sampler [17] samples a large number of trackers from the tracker space dynamically to compensate for target variations. Other approaches have also combined the tracker and detector together [9, 16] but limited the role of the detector so as to assist the tracker as a confidence measurement tool. The benefit of using output from the tracker and detector directly as association candidates for the tracked target, however, has never been fully exploited. We argue that results obtained from the tracker and detector generate redundant association candidates and can complement each other in different scenarios. For example, a drifting tracking result can be corrected by the detection result (Figure 1(a)) and an imprecise detection result can be replaced by a better tracking prediction (Figure 1(b)). In the case of missed detection (Figure 1(c)) and occlusion (Figure 1(d)), the prediction power of the tracker may help to maintain the position and identity of the tracked target. Similarity measurement of appearance model alone is unreliable and the exploration of the interplay among multiple cues in a tracking environment yields promising results [9, 7]. In this paper, we propose a strategy to optimize the association by selecting output of a detector or a tracker at each time step to increase the overall tracking accuracy and precision.

Our method makes the following contributions:

- We present a novel ensemble framework that leverages redundancy and diversity between tracking and detection for robust multi-target tracking. Instead of using detection and classification results to guide the tracker, we treat the tracker and object detector as two independent identities and we keep both their results as association candidates for the tracked target. In each frame, we select the best candidate and assign it to the tracked target (Figure 2). The assignment is scored by a function integrating detection confidence, appearance affinity, and smoothness constraints imposed using geometry and motion information.
- Our approach exploits the discriminative power of the tracker and detector. The weighting factor of each term used in the score function is discriminatively trained. The methodology of data mining for hard negative examples [2] is applied to handle a very large set of artificially generated negative samples.
- In order to deal with unreliable detections, we add additional robustness to our model by mapping the detection confidence into a ranking score by a pre-trained classifier based on multi-scale HOG and texture information.

2 Related work

Building on the success of state-of-the-art object detection methods, object tracking appears to be “easier” to achieve if best matching detection targets can be transitively linked. However, due to the numerous false positives and missed targets in detection results, local data association based on affinity measures between contiguous detections is hard to achieve, hence limiting the ability to find a unique trajectory for each tracked target without drifting [18, 6].

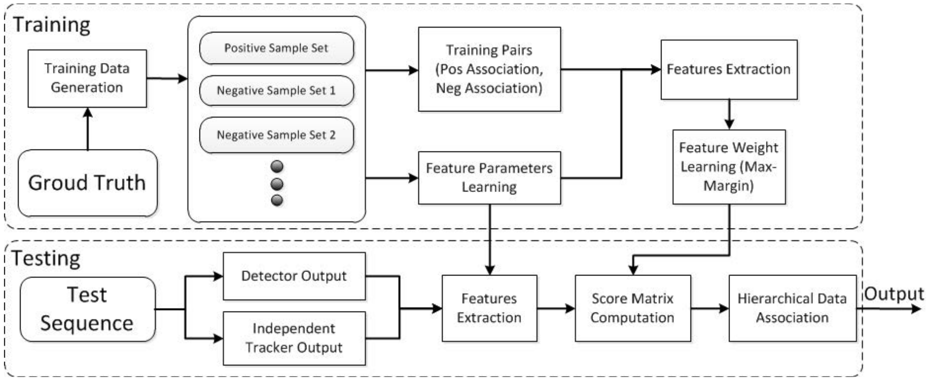


Fig. 3. Framework of our tracking system

On the other hand, global data association tries to solve the problem by optimizing the linkage problem of multiple trajectories simultaneously [13, 14, 11]. Since global methods tend to be computationally expensive, they usually start by detecting short tracklets and iteratively associating them into longer tracks. Leibe *et al.* [19] proposed to couple the output of detector and tracjectory estimation, but trajectories ultimately rely on detections. To overcome the difficulties faced by the global association approaches, Breitenstein *et al.* [9] proposed to deal with the detection uncertainty in a particle filtering framework in which unreliable detection information is complemented by the prediction power of the tracker. In order to increase the association confidence, a boosted classifier is trained online to assess the similarity of tracker-detection pairs. Independent from the detector's output, the classifier term improves the robustness of the tracking result. This coupling framework has also been applied to challenging sport videos [16], which uses a vote-based confidence map to localize people, and the motion model is estimated by the optical flow.

Our approach differs from other data association based tracking methods that perform tracking by associating the output of either the detector or basic tracker only. Unlike these methods, our data association works on results of both a detector and multiple basic trackers.

3 System Overview

Our system is initialized with a human detector and several independent human trackers. Each independent tracker deals with one target. As illustrated in Figure 3, after collecting redundant candidates from outputs of both the detector and independent trackers in testing stage, the hierarchical data association step optimizes the association between tracked targets and candidates. We reduce the association problem to an assignment problem. To manage time complexity, we adopt a greedy-search based association framework using the score matrix between candidates and targets as detailed in Section 4. The score is computed by the dot product between a set of learned weights and features extracted from multiple cues. In addition to color histogram, optical flow, and motion features, we learn an additional target classifier to measure the object detection confidence. Those weights are trained using a max-margin framework, which is designed to give high affinity score for associating candidates with true tracked targets and low score when tracked targets are associated with drifting, false positive candidates or candidates belonging to different targets. To learn the weight parameter, positive samples are obtained from the ground truth and a large number of negative samples are artificially generated to prevent sample selection bias as described in Section 5. We validate our method using publicly available datasets under different challenging conditions and demonstrate superior tracking results that outperform the state-of-the-art algorithms, particularly in terms of accuracy.

4 Ensemble Model

We formulate the multi-object tracking problem as a sum assignment problem that associates tracking candidates obtained from outputs of the tracker and detector to tracked targets of interest. Let $\mathbb{S} = \{s_1, \ldots, s_m\}$ be all tracked targets, where m is the number of objects currently being tracked. Let $\mathbb{R} = \{r_1, \ldots, r_m\}$ indicate the set of all independent trackers. In this paper, a color based particle filter is implemented for each independent tracker $r_i \in \mathbb{R}$. Each particle filter tracker deals with one target and $\mathbb{T} = \{t_1, \ldots, t_m\}$ represents the output of particle filter trackers. The detector's output is denoted as $\mathbb{D} = \{d_1, \ldots, d_n\}$, where n is the number of detection outputs. The set $\mathbb{D} \cup \mathbb{T}$ represents all $m + n$ tracking candidates. The aim of the system is to find the optimal assignment for all tracked targets $\mathbb{S}$ in each frame t, which is measured by the association score of the form:

$$\begin{aligned} &\underset{\{j\}}{\arg\max} \sum_{i=1}^{m} \beta \cdot \varPhi_t(x_i^j) \\ &i \in \mathbb{S}, j \in \mathbb{D} \cup \mathbb{T} \\ &s.t.\ \forall x_a^p, x_b^q \quad p \neq q\ if\ a \neq b, \end{aligned} \tag{1}$$

where x_i^j indicates that the candidate j is assigned to tracked target i, $\{j\}$ denotes a set of selected candidates, β is a vector of model parameters which is learned as presented in Section 5, $\varPhi_t(\cdot)$ represents the association feature set in the current frame, and $\beta \cdot \varPhi_t$ is the score function. The proposed formulation finds optimal links between tracked targets and candidates provided by both the tracker and the detector by assigning at most one candidate to at most one target, and the assignment is evaluated by the affinity score defined in Equation 1. The association feature set $\varPhi = [\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6]$ combines information from different feature spaces, namely the classification confidence, the color histogram, the motion feature and the optical flow feature. Each component is described below in detail.

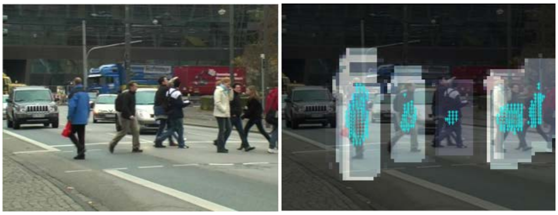


Fig. 4. Visualization of the classification confidence map

Classification Confidence. The classification confidence (ϕ_1) is proportional to the likelihood that the object of interest appears at a given position, and the confidence is derived from the classification result of a binary classifier introduced to gain additional robustness to our discriminative framework [9]. The classifier scores a feature vector x with a dot product function $\omega \cdot x$, where ω is a vector of weighting parameters and x is the feature vector extracted from a given image patch. Weights w are trained using a max-margin framework, the details of which are provided in Section 5. The feature vector x is the concatenation of multi-scale HOG [20] and LBP [21] feature sets. The HOG feature is extracted from a two-level image pyramid. We perform PCA on each HOG feature to manage the curse of dimensionality and improve prediction performance. The cumulative energy threshold for selecting eigenvectors is set at 95%. For image patches with negative score, their classification confidence is set to zero. For those with positive score, the score is normalized by its ranking among scores of positive training samples. For example, an image patch with classification confidence of 0.95 will have higher score than 95% of positive training samples. Figure 4 illustrates the confidence map after applying the binary classifier for human detection. As can be seen, areas of detection targets yield high confidence value.

Color Histogram. The 3D color histogram is built in the Red-Green-Intensity (RGI) space with 5 bins per channel. Given the training pairs, we perform a kernel density estimate for the target and candidate. The similarity between two kernels $g(x_c)$ and $g(x_t)$ is measured by the Bhattacharyya coefficient B, and the likelihood is given by:

$$\phi_2 \propto \exp(-\lambda(1 - B[g(x_c), g(x_t)])) \ , \tag{2}$$

where λ is set to be 5.

Motion Feature. The speed, object scale, and angle represent the motion feature of objects. The speed is modeled by the Normal distribution;
$\phi_3 \propto f_s(\frac{s_t}{s_{t-1}}; \mu^s, \sigma^s)$, where $\frac{s_t}{s_{t-1}}$ is the speed ratio between two frames and f_s is the probability density function. The angle likelihood is modeled by the *von Mises* distribution [22], which is formulated as:

$$\phi_4 = \frac{e^{\kappa^a cos(\theta - \mu^a)}}{2\pi I_0(\kappa^a)} \ , \tag{3}$$

where $I_0(.)$ is the modified Bessel function of order zero. The scale likelihood is modeled by the Normal distribution; $\phi_5 \propto f_l(l; \mu^l, \sigma^l)$, where l is the scale between two frames and f_l is the probability density function. Model parameters $\{\mu^s, \sigma^s, \mu^a, \kappa^a, \mu^l, \sigma^l\}$ are learned from positive training samples.

Optical Flow Feature. The optical flow is precalculated according to [23]. The dominant motion of each region is encoded by a 2D histogram that quantifies both the magnitude and angle of the motion with 10 bins and 8 bins, respectively. The Bhattacharyya coefficient B is used to measure the similarity of histograms $\{H_t, H_c\}$ between the target and candidate. The optical flow score function is given by $\phi_6 \propto \exp(-\tau(1 - B[H_t, H_c]))$, where τ is set to be 5.

4.1 Hierarchical Data Association

Our algorithm employs a hierarchical association strategy to solve Equation 1 by progressively associating outputs of independent trackers and the detector to tracked targets. We use the word "active" to distinguish a target which is not occluded. The association hierarchy consists of three levels. In each level, the assignment is obtained by the Hungarian algorithm. At the first level, it finds the best association between active targets and all candidates. At the second level, the occluded targets are being associated to all unassigned candidates of the detector. If the best association score is below a threshold, the occluded target is moved to the third level, in which, it will be linked to one of the unassigned candidates of the independent trackers based on the association score. The details of the algorithm are described in Algorithm 1.

Greedy Search by Hungarian Algorithm. Given m targets, we solve the assignment problem for detector's n candidates and independent trackers' m candidates through the Hungarian algorithm. The score matrix is defined as:

$$S = \left[\underbrace{\begin{array}{cccc} s_1^1 & s_1^2 & \dots & s_1^n \\ s_2^1 & s_2^2 & \dots & s_2^n \\ \vdots & \vdots & \vdots & \vdots \\ s_m^1 & s_m^2 & \dots & s_m^n \end{array}}_{\text{Detector's Candidates}} \middle| \underbrace{\begin{array}{cccc} s_1^{n+1} & -\infty & \dots & -\infty \\ -\infty & s_2^{n+2} & \dots & -\infty \\ \vdots & \vdots & \vdots & \vdots \\ -\infty & -\infty & \dots & s_m^{n+m} \end{array}}_{\text{Trackers' Candidates}} \right]$$

The score in S is computed by $s_i^j = \beta \cdot \Phi(x_i^j)$ where i and j are row and column, respectively. The negative infinity value of the off-diagonal components represents self-association rule of the tracking result [14] as it is designed to be linked to one specific target only.

Occlusion Handling. We set the "enter" and "exit" regions along image borders after the first two frames in a typical surveillance setting similar to [9]. If the best association score for a target is below a threshold in the first level association, it will be marked as "occluded". In addition, an assigned target in the first level with a lower classification confidence score than the threshold will also be marked as "occluded". We activate an occluded target only if its associated candidate returns a classification confidence score greater than the threshold. For an occluded target, the association feature set is not updated until it becomes active again. An occluded target will be deleted if it stays in the "exit" region for more than 5 frames or remains "occluded" for more than 20 frames in the "non-exit" regions. Deletion of an active target depends on its position with respect to the "exit" region.

Algorithm 1: Association Framework

```
Input: Targets $\{1, \ldots, i, \ldots, m\}$, Candidates $\{1, \ldots, j, \ldots, (n+m)\}$.
Output: Association Results.
Compute the association feature for all active targets and candidates;
Compute the score $s_i^j = \beta \cdot \Phi(x_i^j)$ and enter it to the score matrix;
Apply the Hungarian algorithm to solve the assignment problem;
for each assignment $(i, j)$ do
    if $s_i^j <$threshold or $\phi_1(j) <$threshold then
        invalidate the assignment;
    end
end
Recompute the score matrix for all occluded targets and unassigned
  detection candidates;
Apply the Hungarian algorithm to solve the assignment problem;
for each assignment $(i, j)$ do
    if $s_i^j <$threshold or $\phi_1(j) <$threshold then
        invalidate the assignment;
    end
end
if active target is assigned then
    update the feature set;
else
    active target is set as occluded;
end
if occluded target is assigned then
    occluded target is set as active and its feature set is updated;
end
For all unassigned occluded targets, associate them to corresponding
  tracking result.
```

4.2 Detector and Independent Trackers

We use a state-of-the-art deformable part based detector to detect the occurrences of human targets in every frame [2]. The tracker we use is similar to [24] where the particle's observation model is built upon the RGI color histogram [9]. We keep track of two frames for updating the observation template: one is the first frame that the target appears in and the other is the latest frame in which the target is in the "active" state. A new independent tracker is initialized for a target that has higher classification confidence value than the threshold in two continuous frames and has no existing tracker associated to it.

5 Discriminative Learning

The model parameters β in Equation 1 is learned discriminatively. Consider a training set $\{(y_i, \Phi(o_i))\}_{i=1}^N$, where $y_i \in \{+1, -1\}$ is the label and $\Phi(o_i) \in \mathbb{R}^n$ is the feature vector extracted from the training instance o_i that includes an assigned candidate. The objective is to learn a model that assigns the score to the instance with a function of the form $\beta \cdot \Phi(\cdot)$, where β is a vector of model parameters. The formulation, in analogy to classical SVMs, leads to the following optimization problem:

$$f(\beta) = \min_{\beta} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{N} \ell(\beta; (\Phi(o_i), y_i)) \ , \tag{4}$$

where $\ell(\beta; (\Phi(o_i), y_i)) = \max\{0, 1 - y_i \langle \beta, \Phi(o_i) \rangle\}$ is the hinge loss function and the constant C is chosen experimentally as the weight for the penalty. Stochastic sub-gradient method [2, 25] is applied for solving this problem.

Typically positive samples are given and we manually generate the "hard" negative samples. For training, negative samples are randomly generated for three different kinds of scenarios: tracking drift, false positive and mismatch. As shown in Figure 5(b), samples for tracking drift are picked as image patches that have between 0% $\sim$ 25% overlap with the ground truth. False positive accounts for cases where there is no overlap between tracked objects in the ground truth and candidates obtained from the tracker and detector (Figure 5(c)). Mismatch represents identity switch, in which a tracked target is connected to a wrong candidate (Figure 5(d)). Compared with the number of positive samples, the number of negative instances is very large. To deal with a large set of samples, we apply the data-mining algorithm proposed in [2] for training our model efficiently.

6 Experiments

6.1 Datasets

We evaluate our tracking algorithm on four public challenging datasets: TUD Crossing, TUD Campus, ETHZ Central and UBC Hockey [9]. The video in the UBC Hockey dataset is acquired by a moving camera and static cameras are used for videos in other datasets. These four datasets present a wide range of challenges due to heavy inter-person occlusion, poor image quality, and low image contrast between targets and the background. Videos in these datasets also cover different viewpoints and capture various types of movements. In all experiments, we define "entry" and "exit" zones manually for each sequence and no other scene

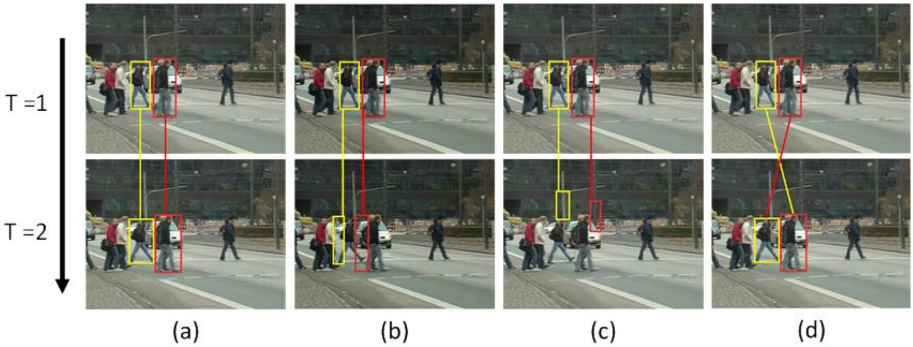


Fig. 5. Positive and negative samples generated for the tracking problem: (a) correct tracking, (b) tracking drift, (c) false positive, (d) mismatch

or calibration knowledge is leveraged. We employ the discriminatively trained deformable parts model [2] as the human detector. The detector uses publicly available and pre-trained model for TUD Crossing and TUD Campus datasets. The deformable parts model is re-trained for ETHZ Central and UBC Hockey datasets to boost the detection rate as the quality of images is much poorer in these videos. None of the video frames in the dectector training are used for testing.

6.2 Parameter Training

We obtained the model parameter β as described in Section 5. We use the same parameter for TUD Crossing and TUD Campus datasets, which is trained on the first 25 frames of the TUD Crossing video. The parameter β is trained for ETHZ Central by using the first video sequence in the dataset and our tracking algorithm is tested on the second video sequence. For UBC Hockey, the first 25 frames are used for training. None of the video frames in the parameter training set are used for testing.

6.3 Quantitative Evaluation

We adopt CLEAR MOT metrics [26] to evaluate the tracking performance of our algorithm. Key measurements of the metrics include: precision score (MOTP) measured by intersection over union of bounding boxes, and an accuracy score (MOTA) which is composed of false negative rate (FN), false positive rate (FP), and number of ID switches (ID Sw.). Results of our algorithm are reported in Table 1 (shown in top row) after conducting experiments on aforementioned four datasets. In general, the results indicate that our approach achieves high tracking accuracy with very few number of ID switches. In our experiments, false positives usually are caused by the drift of occluded targets since it is hard to update the motion model in time during occlusion. For example, the rapid change of movements in the UBC Hockey video increases the chance of false positives. The failure of the human detector is the main reason for false negatives in the result since several persons are not detected and corresponding independent trackers are not initialized in the video. A typical detection failure happens when occlusions persist over several video frames. For example, as shown in Figure 6, one of the persons sitting in the lower-right corner is never detected. Occlusion is also the culprit for ID switches. If a newly detected person bears similar appearance with an occluded target, the ID of the occluded one may be mistakenly assigned to another target.

For comparison, we list the results of three competing approaches for these sequences: (i) On-line Multi-Person Tracking-by-Detection [9] on TUD Crossing, TUD Campus, ETHZ Central and UBC Hockey; (ii) Coupled detection and trajectory estimation [19] on ETHZ Central; (iii) Boosted particle filter [27] on UBC Hockey. As shown in Table 1, we outperform the competing approaches on all datasets in terms of tracking accuracy. As for the tracking precision, our results are comparable with the best reported performance measures.

Table 1. CLEAR MOT evaluation results on four datasets. Our results are in the top row for each dataset. The best results are in bold.

Dataset	MOTP(%)	MOTA(%)	FP(%)	FN(%)	ID Sw.
TUD Crossing	70.77	**89.38**	**1.09**	**9.33**	**2**
TUD Crossing [9]	**71.00**	84.30	1.40	14.10	**2**
TUD Campus	**67.76**	**84.82**	**0.00**	**15.18**	**0**
TUD Campus [9]	67.00	73.30	0.10	26.40	2
ETHZ Central	**71.49**	**75.40**	0.36	**24.24**	**0**
ETHZ Central [9]	70.00	72.90	**0.30**	26.8	**0**
ETHZ Central [19]	66.00	33.80	14.70	51.30	5
UBC Hockey	**71.61**	**91.75**	1.76	**6.49**	**0**
UBC Hockey [9]	57.00	76.50	1.20	22.30	**0**
UBC Hockey [27]	51.00	67.80	**0.00**	31.30	11

To fully evaluate the benefit of the ensemble tracking-by-detection framework proposed in this paper, we also present the performance of component-wise analysis. The default method used the output of both the part-based detector and independent particle filter trackers to accomplish data association. Variant (a) leverages output of particle filter trackers alone as tracking candidates while variant (b) leverages output of only the part-based detector. As shown in Table 2, the default method performs better in term of accuracy, false positive, false negative and ID switches over the result of variants due to the optimal selection of output of both components. The lower precision score of the default method in three of four datasets is related to the way MOTP is computed. Since MOTP only measures the positional deviation of detected targets from their ground truth, an increase in the number of detected targets can lead to lower overall precision. This is the case since the default method is able to track a greater number of targets than either of the variants.

Table 2. CLEAR MOT evaluation results on component-wise evaluation of our approach. Variant (a) leverages output of the independent trackers only. Variant (b) leverages output the detector only. The best results are in bold.

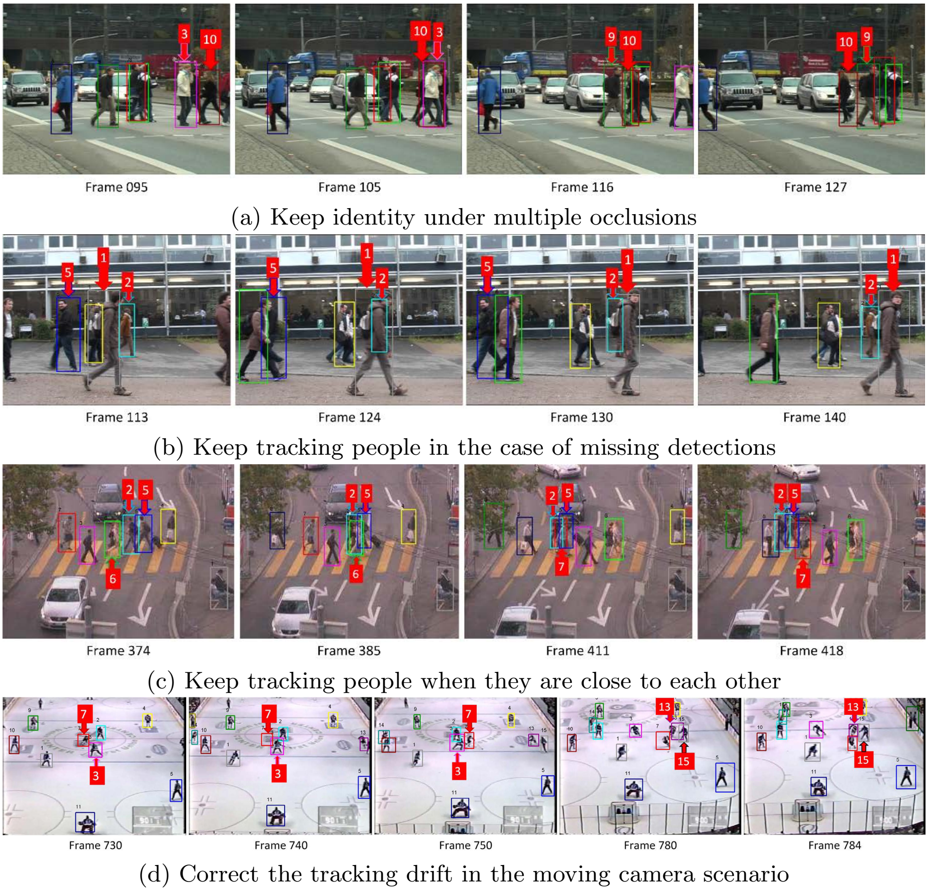
Dataset	MOTP(%)	MOTA(%)	FP(%)	FN(%)	ID Sw.
TUD Crossing (**Default**)	**70.77**	**89.38**	**1.09**	**9.33**	**2**
TUD Crossing (a)	63.06	58.13	20.04	21.63	19
TUD Crossing (b)	69.46	82.14	6.85	10.81	36
TUD Campus (**Default**)	67.76	**84.82**	**0.00**	**15.18**	**0**
TUD Campus (a)	62.80	47.19	18.15	33.99	**0**
TUD Campus (b)	**70.08**	59.74	17.16	22.44	3
ETHZ Central (**Default**)	71.49	**75.4**	**0.36**	**24.24**	**0**
ETHZ Central (a)	59.25	26.74	37.97	34.94	23
ETHZ Central (b)	**75.59**	72.91	1.96	24.78	7
UBC Hockey (**Default**)	71.61	**91.75**	**1.76**	**6.49**	**0**
UBC Hockey (a)	58.32	80.41	4.85	14.56	26
UBC Hockey (b)	**73.42**	82.84	3.41	13.57	10

6.4 Qualitative Evaluation

Figure 6 shows our qualitative evaluation. The first row presents the ability of our approach to keep the identity for target #10 even when the target has been occluded by multiple targets in the sequence. The scenario in the second row shows that we can keep updating the location of occluded targets by using the tracker's prediction in the case of missing detections. The third row demonstrates that our tracker can differentiate targets well when they are very close to each other. The last row shows a sequence shot from a moving camera. Although the motion model loses its accuracy due to abrupt changes of movements, our tracker can correct tracking drifts by switching to associate detection results to tracked targets.

7 Conclusion

We have presented a novel ensemble framework based on the tracking-by-detection approach. Association candidates in this integrated model come from independent trackers and object detector. The best candidate is selected based on a score function that integrates classification confidence, appearance affinity, and smoothness constraints imposed using geometry and motion information. Model parameters of the score function are discriminatively trained. In order to improve the detection confidence in complex scenes, the framework incorporates an additional target classifier that is also trained discriminatively. As our experiments show, the proposed approach achieves good performance in different datasets with a variety of scenarios and outperforms other state-of-the-art methods. Finally, the performance of the algorithm could be improved if we enhance the discriminative model for visual matching in the tracker by on-line metric learning.



(a) Keep identity under multiple occlusions

(b) Keep tracking people in the case of missing detections

(c) Keep tracking people when they are close to each other

(d) Correct the tracking drift in the moving camera scenario

Fig. 6. Tracking results of our approach on TUD Crossing, TUD Campus, ETHZ Central and UBC Hockey datasets

References

1. Yang, H., Shao, L., Zheng, F., Wang, L., Song, Z.: Recent advances and trends in visual tracking: A review. Neurocomputing 74, 3823–3831 (2011)
2. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI 32, 1627–1645 (2010)
3. Wang, X., Han, T.X., Yan, S.: An HOG-LBP human detector with partial occlusion handling. In: CVPR (2009)
4. Vijayanarasimhan, S., Grauman, K.: Large-scale live active learning: Training object detectors with crawled data and crowds. In: CVPR (2011)
5. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: CVPR (2008)
6. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised On-Line Boosting for Robust Tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)

7. Choi, W., Savarese, S.: Multiple Target Tracking in World Coordinate with Single, Minimally Calibrated Camera. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 553–567. Springer, Heidelberg (2010)
8. Stalder, S., Grabner, H., Gool, L.V.: Cascaded Confidence Filtering for Improved Tracking-by-Detection. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 369–382. Springer, Heidelberg (2010)
9. Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Online multiperson tracking-by-detection from a single, uncalibrated camera. PAMI 33, 1820–1833 (2011)
10. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning. PAMI 33, 1619–1632 (2011)
11. Andriyenko, A., Schindler, K.: Multi-target tracking by continuous energy minimization. In: CVPR (2011)
12. Kuo, C.H., Nevatia, R.: How does person identity recognition help multi-person tracking? In: CVPR (2011)
13. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: CVPR (2008)
14. Huang, C., Wu, B., Nevatia, R.: Robust Object Tracking by Hierarchical Association of Detection Responses. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 788–801. Springer, Heidelberg (2008)
15. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: CVPR (2011)
16. Yao, A., Uebersax, D., Gall, J., Gool, L.V.: Tracking People in Broadcast Sports. In: Goesele, M., Roth, S., Kuijper, A., Schiele, B., Schindler, K. (eds.) Pattern Recognition. LNCS, vol. 6376, pp. 151–161. Springer, Heidelberg (2010)
17. Kwon, J., Lee, K.M.: Tracking by sampling trackers. In: ICCV (2011)
18. Wu, B., Nevatia, R.: Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. IJCV 82, 184–204 (2007)
19. Leibe, B., Schindler, K., Gool, L.V.: Coupled detection and trajectory estimation for multi-object tracking. In: ICCV (2007)
20. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
21. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. PAMI 28, 2037–2041 (2006)
22. Song, B., Jeng, T.Y., Staudt, E., Roy-Chowdhury, A.K.: A Stochastic Graph Evolution Framework for Robust Multi-target Tracking. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 605–619. Springer, Heidelberg (2010)
23. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. PAMI 33, 500–513 (2011)
24. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-Based Probabilistic Tracking. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part I. LNCS, vol. 2350, pp. 661–675. Springer, Heidelberg (2002)
25. Shalev-Shwartz, S., Singer, Y., Srebro, N.: Pegasos: Primal estimated sub-gradient solver for SVM. In: ICML (2007)
26. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the CLEAR MOT metrics. Image Video Processing 1, 1–10 (2008)
27. Okuma, K., Taleghani, A., Freitas, N.: A Boosted Particle Filter: Multitarget Detection and Tracking. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 28–39. Springer, Heidelberg (2004)