ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A New Novel Approach for Online Offline Multi Object Tracking with Different Complex Scenarios

Mandapuram Venkatesh
PG Scholar (DECE)
Department of ECE
Tudi Narasimha Reddy Institute of Technology & Sciences,
Hydearabad.

Mr.Sathish Parvatham
Associate Professor
Department of ECE
Tudi Narasimha Reddy Institute of Technology & Sciences,
Hydearabad.

Dr. Samalla Krishna
Professor
Department of ECE
Tudi Narasimha Reddy Institute of Technology & Sciences,
Hydearabad.

ABSTRACT

In this paper, we address the problem of automatically detecting and tracking a variable number of persons in complex scenes using a monocular, potentially moving, uncalibrated camera. We propose a novel approach for multi-person tracking-by detection in a particle filtering framework. In addition to final high-confidence detections, our algorithm uses the continuous confidence of pedestrian detectors and online trained, instance-specific classifiers as a graded observation model. Thus, generic object category knowledge is complemented by instance-specific information. The main contribution of this paper is to explore how these unreliable information sources can be used for robust multi-person tracking.

The algorithm detects and tracks a large number of dynamically moving persons in complex scenes with occlusions, does not rely on background modeling, requires no camera or ground plane calibration, and only makes use of information from the past. Hence, it imposes very few restrictions and is suitable for online applications. Our experiments show that the method yields good tracking performance in a large variety of highly dynamic scenarios, such as typical surveillance videos, webcam footage, or sports sequences. We demonstrate that our algorithm outperforms other methods that rely on additional information. Furthermore, we analyze the influence of different algorithm components on the robustness

INTRODUCTION

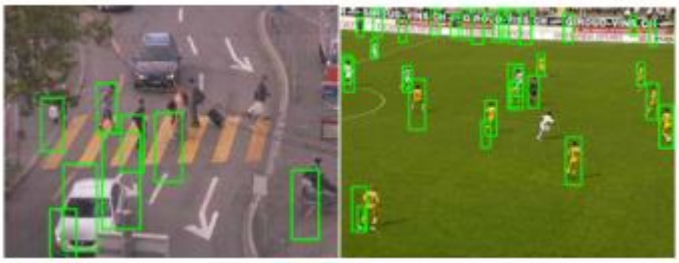
New video cameras are installed daily all around the world, as webcams, for surveillance, or for a multitude of other purposes. As this happens, it becomes increasingly important to develop methods that process such data streams automatically and in real-time, reducing the manual effort that is still required for video analysis. Of particular interest for many applications is the behavior of persons, e.g., for traffic safety, surveillance, or sports analysis. As most tasks at semantically higher levels are based on trajectory information, it is crucial to robustly detect and track people in dynamic and complex real-world scenes. However, most existing multiperson tracking methods are still limited to special application scenarios. They require either multi-camera input, scenespecific knowledge, a static background, or depth information, or are not suitable for online processing. In this paper, we address the problem of automatically detecting and tracking a variable number of targets in complex scenes from a single, potentially moving, uncalibrated camera, using a causal (or online) approach. This problem is very challenging, because there are many sources of uncertainty for the object locations such as measurement noise, clutter, changing background, and significant occlusions final non-maximum suppression stage. This said, it is not guaranteed that the shape of the confidence volume in-between those locations will support tracking. In particular, a majority of the densities' local maxima correspond to false positives that may deteriorate the tracking results, especially during occlusions and when several interacting targets are present. The main contribution of our work is the exploration how this unreliable information source can be used for robust multi-person tracking. Our algorithm achieves this robustness through a careful interplay between object detection, classifi- cation, and target tracking components. Typically, a bottom-up process deals with target representation and localization, trying to cope with changes in the appearance of the tracked targets, and a top-down process performs data association and filtering to deal with object dynamics. Correspondingly, our approach is based on a combination of a general, class-specific pedestrian detector to localize people and a particle filter to predict the target locations, incorporating a motion model. To complement the generic object category knowledge from the detector, our algorithm trains person-specific classifiers during run-time to distinguish between the tracking targets. This paper makes the following contributions: 1) We combine a generic class-specific object detector and particle filtering for robust multi-person tracking suitable for online applications. following section, Section 3 describes the algorithm and several important design choices. Section 4 presents a quantitative evaluation on a large variety of datasets and a comparison to other algorithms. In Section 5, the robustness of the observation model is discussed in detail. Section 6 concludes the paper with a summary and outlook

Particle Filtering

Particle filters were introduced to the vision community to estimate the multi-modal distribution of a target's state space [19]. Other researchers extended the framework for multiple targets by either representing all targets jointly in a particle filter [43] or by extending the state space of each target to include components of other targets [41]. In the first approach, a fixed number of particles represent a varying number of targets. Hence, new targets have to "steal" particles from existing trackers, reducing the accuracy of the approximation. In the second approach, the state space becomes increasingly large, which may require a very large number of particles for a good representation. Thus, the computational complexity increases exponentially with the number of targets. To overcome these problems, most methods employ one particle filter per target using a small state space and deal with interacting targets separately [21], [24], [38]. Tracking-by-Detection. While many tracking methods rely on background subtraction from one or several static cameras [3], [20], [24], [42], [49], recent progress in object detection has stimulated the interest in combining tracking and detection.

DETECTOR CONFIDENCE PARTICLE FILTER

For many tracking applications, only past observations can be used at a certain time step to estimate the location of objects. Within this context, Bayesian Sequential Estimation is a popular approach, which recursively estimates the time-evolving posterior distribution of the target locations conditioned on all observations seen so far. This filtering distribution can be approximated by Sequential Monte Carlo Estimation (or Particle Filtering), which represents the distribution with a set of weighted particles and consists of a dynamic model for prediction and an observation model to evaluate the likelihood of a predicted state [10].

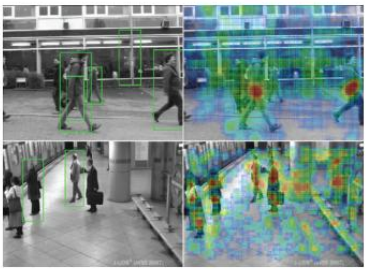


The output of a person detector (right: ISM [25], left: HOG [9]) with false positives and missing detections.

Algorithm Overview Our algorithm implements a first-order Markov model, considering only information from the current and the last time step, and integrates both class-specific and target-specific information in the observation model. A separate particle filter (tracker) is automatically initialized for each person detected with high confidence. To achieve the necessary robustness, the information from an object detector is integrated in two ways. First, the algorithm carefully assesses the high-confidence detections in each frame and selects maximally one to track one particular target.

Detector Confidence.

At the core of our approach lies the confidence density built up by person detectors in some form. This is the case for both sliding-window based detectors such as HOG [9] and for feature-based detectors such as ISM [25]. In the sliding-window case, this density is implicitly sampled in a discrete 3D grid (location and scale) by evaluating the different detection windows with a classifier. In the ISM case, it is explicitly created in a bottom-up fashion through probabilistic votes cast by matching, local features. In order to arrive at individual detections, both types of approaches search for local maxima in the density volume and then apply some



Detector output (top: ISM [25], bottom: HOG [9]), showing high-confidence detections (left, green rectangles) and the detector confidence (right, shaded overlay). The confi- dence density often contains useful information at the location of missing detections, which we exploit for tracking.

Particle Filtering Our tracking algorithm is based on estimating the distribution of each target state by a particle filter. The state x = {x, y, u, v} consists of the 2D image position (x, y) and the velocity components (u, v). We employ the bootstrap filter, where the state transition density (or prior kernel) is used as importance distribution to approximate the probability density function [16]. The importance weight w i t for each particle i at time step t is described by:

$$w_t^i \quad \propto \quad w_{t-1}^i \cdot p(o_t|x_t^i).$$

Since re-sampling is carried out in each time step using a fixed number of N = 100 particles, w i t−1 = 1 N is a constant and can be ignored. Thus, Eq. (1) reduces to the likelihood of a new observation ot given the propagated particles x i t , which we estimate as described in Sec. 3.4 (Eq. (6)).

Size And Position.

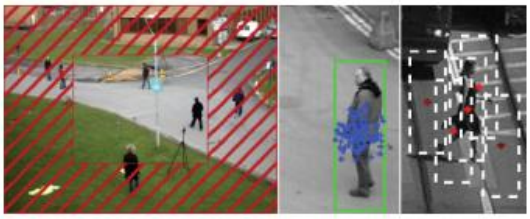
Instead of including the size of the target in the state space of the particles, the target size is set to the average of the last four associated detections. In our experiments, this yielded better results, possibly because the number of particles necessary to estimate a larger state space is growing exponentially. Although represented by a (possibly multi-modal) distribution, a single position of the tracking target at the current time step is sometimes required (e.g., for visualization or evaluation).

Motion Model.

To propagate the particles, we use a constant velocity motion model

$$\begin{aligned}(x, y)_t &= (x, y)_{t-1} + (u, v)_{t-1} \cdot \Delta t + \varepsilon_{(x,y)} \\ (u, v)_t &= (u, v)_{t-1} + \varepsilon_{(u,v)}.\end{aligned}$$

The process noise ε(x,y) , ε(u,v) for each state variable is independently drawn from zero-mean normal distributions. The initial variances σ 2 (x,y) and σ 2 (u,v) for position and velocity noise are set proportionally to the size of the tracking target

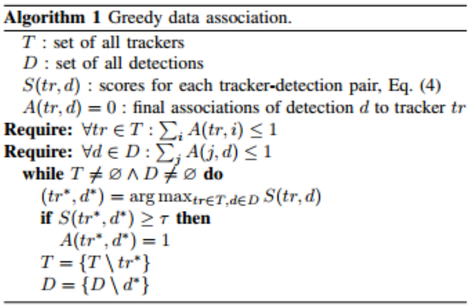


ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The initialization and termination region for a typical surveillance scenario (left). The initial particles are drawn from a normal distribution centered at the detection (middle). The weight of each particle is determined by evaluating the respective image patch (right). During tracking, they decrease inversely proportional to the number of successfully tracked frames (down to a lower limit). Hence, the longer a target is tracked successfully, the less the particles are spread.

Initialization and Termination.

Object detection yields fully automatic initialization. The algorithm initializes a new tracker for an object that has subsequent detections with overlapping bounding boxes, which are neither occluded nor associated to an already existing tracker. In order to avoid persistent false positives from similar looking background structures (such as windows, doors, or trees), we only initialize trackers from detections that appear in a zone along the image borders for sequences where this is reasonable, such as for typical surveillance settings.



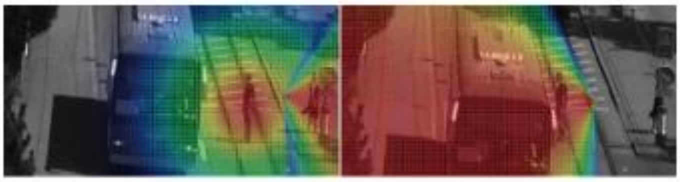
```
Algorithm 1 Greedy data association.
    $T$ : set of all trackers
    $D$ : set of all detections
    $S(tr,d)$ : scores for each tracker-detection pair, Eq. (4)
    $A(tr,d) = 0$ : final associations of detection $d$ to tracker $tr$
Require: $\forall tr \in T : \sum_i A(tr,i) \leq 1$
Require: $\forall d \in D : \sum_j A(j,d) \leq 1$
    while $T \neq \varnothing \wedge D \neq \varnothing$ do
        $(tr^*, d^*) = \arg\max_{tr \in T, d \in D} S(tr,d)$
        if $S(tr^*, d^*) \geq \tau$ then
            $A(tr^*, d^*) = 1$
        $T = \{T \setminus tr^*\}$
        $D = \{D \setminus d^*\}$
```

Data Association In order to decide which detection should guide which tracker, we solve a data association problem, assigning at most one detection to at most one target. The optimal single-frame assignment can be obtained by the Hungarian algorithm [22]. In our experiments, we however found that a greedy algorithm achieves similar results at lower computational cost. Greedy Data Association. The matching algorithm works as follows (see Algorithm 1): First, a matching score matrix S for each pair (tr, d) of tracker tr and detection d is computed as described below. Then, the pair (tr∗ , d∗) with maximum score is iteratively selected, and the rows and columns belonging to tracker tr and detection d in S are deleted.

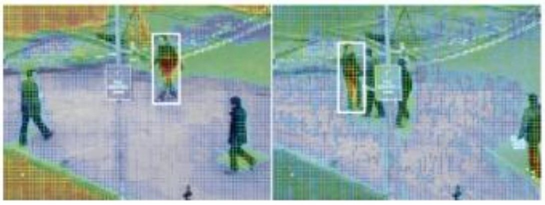
$$S(tr,d) = g(tr,d) \cdot (c_{tr}(d) + \alpha \cdot \sum_{p \in tr}^{N} p_{\mathcal{N}}(d-p)),$$

where pN (d−p) ∼ N (posd −posp; 0, σ2) denotes the normal distribution evaluated for the distance between the position of detection d and a particle p, and g(tr, d) is a gating function described next. The last term of (Eq. (4)) measures the density of the particle distribution, rewarding associations where the particles are densely distributed around the detection. Gating Function. Not only the distance of a detection to the tracker is important, but also its location with respect to the motion direction. Therefore, a gating function g(tr, d) additionally assesses each detection. It consists of the product of two factors:

$$g(tr,d) = p(size_d|tr)p(pos_d|tr)$$



The gating function depends on the velocity of the target, resulting in different 2D cone angles or a radial decay



The classifier response (heat map) visualized for one tracking target (white). As the classifier is adapted continuously, it becomes more discriminative (right: 20 frames later)

$$= \begin{cases} p_{\mathcal{N}}(\frac{size_{tr} - size_d}{size_{tr}}) \cdot p_{\mathcal{N}}(|d - tr|) & \text{if } |v_{tr}| < \tau_v \\ p_{\mathcal{N}}(\frac{size_{tr} - size_d}{size_{tr}}) \cdot p_{\mathcal{N}}(dist(d, v_{tr})) & \text{otherwise.} \end{cases}$$

Observation Model To compute the weight wtr,p for a particle p of the tracker tr, our algorithm estimates the likelihood of a particle. For this purpose, we combine different sources of information, namely the associated detection d ∗ , the intermediate output of the detection algorithm, and the output of the classifier ctr

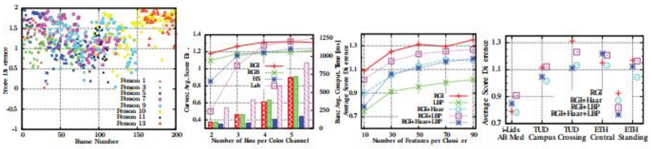
$$w_{tr,p} = \underbrace{\beta \cdot \mathcal{I}(tr) \cdot p_{\mathcal{N}}(p - d^{*})}_{\text{detection}} + \underbrace{\gamma \cdot d_c(p) \cdot p_o(tr)}_{\text{det. confidence}} + \underbrace{\eta \cdot c_{tr}(p)}_{\text{classifier}}$$

EXPERIMENTS

Datasets There is no generally accepted benchmark available for multiperson tracking. Therefore, most related publications have carried out experiments on their own sequences, which we have tried to combine. Thus, we evaluate on a large variety of challenging sequences: ETHZ Central [26], TUD Campus and TUD Crossing [1], i-Lids AB [18], [45], UBC Hockey [7], [33], PETS’09 S2.L1–S2.L3 [12], ETHZ Standing [14], and our own Soccer dataset.2 These sequences are taken from both static and moving cameras, and they vary with respect to viewpoint, type of movement, and amount of occlusion. While some datasets show rather classical surveillance and security scenarios from an elevated viewpoint, others are captured at eye level and are typical for robot / car navigation and traffic safety applications, while some are sports sequences with abrupt motion changes of the players and moving cameras.

Qualitative Analysis ETHZ Central.

The output of the ISM detector is very noisy for the ETHZ Central dataset (Fig. 8, top). The cars and road markings produce many false positives, and pedestrians are often not detected. Only a few detections consistently match the targets throughout the sequence (e.g., the blue tracker in Fig. 8, bottom, gets assigned a detection only every 30 frames). Thus, the trackers often rely on the detector and classifier confidence. Furthermore, there are many occlusions, e.g., when people walk in parallel. Hence, the correct association of detections to trackers is a key factor of our algorithm. TUD Campus. The ISM detections are more accurate for the TUD Campus dataset. On average, a tracker is associated for the Soccer dataset)

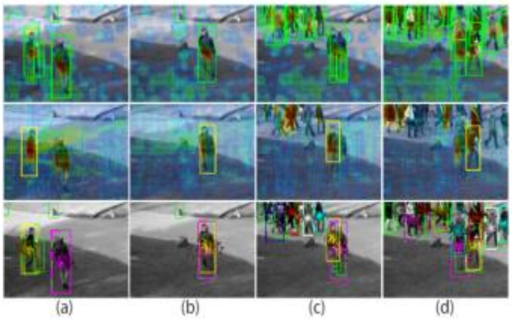


Classifier evaluation on the TUD Crossing sequence with 50 RGI features and 3 bins per color channel. We plot the difference between the classifier score on the correct target and the highest score on all other targets. (b) Evaluation of performance (left scale) and computation time (bars, right scale) for different color features. (c) Evaluation of the number of features per classifier. (d) Evaluation of feature combinations for some datasets.

Result for the ETHZ Central dataset (top: final ISM detections (green) and detector confidence (heat map)), tracking result (bottom)

Quantitative Analysis We use the CLEAR MOT metrics [4] to evaluate the tracking performance. This returns a precision score MOTP (intersection over union of bounding boxes) and an accuracy score MOTA (composed of false negative rate, false positive rate, and number of identity switches).

Dataset	MOTP	MOTA	FN	FP	ID Sw.
ETHZ Central	70.0%	**72.9%**	26.8%	0.3%	0
Leibe *et al.* [26]	66.0%	33.8%	51.3%	14.7%	5
UBC Hockey	57.0%	**76.5%**	22.3%	1.2%	0
Okuma *et al.* [33]	51.0%	67.8%	31.3%	0.0%	11
i-Lids Easy	67.0%	**78.1%**	16.4%	5.3%	18
i-Lids Medium^{*3}	66.0%	**76.0%**	22.0%	2.0%	2
Huang *et al.* [18]	-	68.4%	29.0%	13.7%	-
Wu and Nevatia [45]	-	55.3%	37.0%	22.8%	-
TUD Campus	67.0%	**73.3%**	26.4%	0.1%	2
TUD Crossing	71.0%	**84.3%**	14.1%	1.4%	2
Soccer	67.0%	**85.7%**	7.9%	6.2%	4
PETS’09 S2.L1	56.3%	**79.7%**	-	-	-
PETS’09 S2.L1^{*4}	56.7%	74.9%	-	-	-
Yang *et al.* [47]	53.8%	75.9%	-	-	-
Berclaz *et al.* [3]^{5}	ca. 60%	ca. 66%	-	-	-
PETS’09 S2.L2	51.3%	**50.0%**	-	-	-
PETS’09 S2.L3	52.1%	**67.5%**	-	-	-



Visualization of detector output (top), classifier output for the yellow target (middle), and particle filter output (bottom; dashed bounding boxes are detections associated to the tracker with the respective color)

CONCLUSION

We have presented a novel method for online multi-object tracking-by-detection, exploring the capabilities of an approach that relies only on 2D image information from one single, uncalibrated camera, without any additional scene knowledge. The main challenge for tracking algorithms are unreliable measurements, i.e., in the case of tracking-by-detection, false positives and missing detections. The contribution of our work is thus to explore how this unreliable information source can be used for robust multi-person tracking. The key factors of our algorithm are: (1) careful selection and association of final detections using target-specific classifiers trained during run-time, (2) utilization of the continuous output of detector and classifier, and (3) robust combination of unreliable information for multi-person tracking using particle filtering. While the data association algorithm handles false positive detections, different observation model terms help overcome problems with missing detections. They are complementary, as they are trained on different features and training data.

While instance-specific information is beneficial to resolve ambiguous situations between different targets, class-specific knowledge helps differentiate between object and background. For this purpose, the detector confidence term guides the particles of the filter primarily when no discrete highconfidence detection is issued by the detector. Although this is beneficial for situations with missing detections, it can also misguide trackers to image areas with high confidence on background structures. On the other hand, the classifier term helps localize particles more accurately, adapting online to the appearance of the targets. However, the classifier requires some amount of training data to work reliably and hence does neither help in situations shortly after initialization nor if the appearance of a target changes heavily during occlusions. Our experiments have shown that the method achieves a good performance on a large variety of application scenarios outperforming other state-of-the-art algorithms, some of which rely on scene-specific information, multiple calibrated cameras, or global optimization. To increase the robustness during partial occlusions, a part-based detector would be beneficial. Also, the detector could be trained for specific applications and the motion model could be specialized, e.g., for applications in sports television broadcasting. Furthermore, if applied to a specific scenario, scene-specific information could be used to help resolve ambiguities, restricting motion to a ground plane or providing information about obstacles. Finally, the method could be enhanced by taking advantage of a more sophisticated estimation framework than particle filtering.

REFERENCES

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In IEEE Comp. Vision and Pattern Rec., 2008.

[2] S. Avidan. Ensemble tracking. IEEE T. Pattern Anal. and Machine Intell., 29(2):261–271, 2007.

[3] J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In IEEE Comp. Vision and Pattern Rec., 2006.

[4] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR

MOT metrics. J. Image and Video Processing, (3):1–10, 2008.

[5] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Markovian tracking-by-detection from a single, uncalibrated camera. In IEEE Workshop Performance Evaluation of Tracking and Surveillance, 2009.

[6] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In IEEE Int. Conf. Comp. Vision, 2009.

[7] Y. Cai, N. de Freitas, and J. J. Little. Robust visual tracking for multiple targets. In Eur. Conf. Comp. Vision, 2006.