Multi-target Tracking in Crowded Scenes

Jie Yu[1], Dirk Farin[1], and Bernt Schiele[2]

[1] Corporate Research Advance Engineering Multimedia
Robert Bosch GmbH, Germany
[2] MPI Informatics, Saabrucken, Germany

Abstract. In this paper, we propose a two-phase tracking algorithm for multi-target tracking in crowded scenes. The first phase extracts an over-complete set of tracklets as potential fragments of true object tracks by considering the local temporal context of dense detection-scores. The second phase employs a Bayesian formulation to find the most probable set of tracks in a range of frames. A major difference to previous algorithms is that tracklet confidences are not directly used during track generation in the second phase. This decreases the influence of those effects, which are difficult to model during detection (e.g. occlusions, bad illumination), in the track generation. Instead, the algorithm starts with a detection-confidence model derived from a trained detector. Then, tracking-by-detection (TBD) is applied on the confidence volume over several frames to generate tracklets which are considered as enhanced detections. As our experiments show, detection performance of the tracklet detections significantly outperforms the raw detections. The second phase of the algorithm employs a new multi-frame Bayesian formulation that estimates the number of tracks as well as their location with an MCMC process. Experimental results indicate that our approach outperforms the state-of-the-art in crowded scenes.

1 Introduction

Tracking is a key issue in various video-based applications, such as surveillance, video retrieval systems, robotics, etc. However, multi-target tracking, especially in crowded scenes, is still one of the most challenging problems, due to difficulties such as occlusions, association complexity and measurement noise.

In recent years, much progress in object detection has been made and many detector-based tracking approaches e.g. [1,10,18] have been proposed. In contrast to background-model based approaches, detector-based tracking is robust against changing backgrounds and moving cameras. It can also be used in crowded scenes, where learning and updating the background model are often impractical.

However, object-model learning is also challenging. For some object classes with large appearance variations, e.g. people, detectors can not distinguish objects from the background reliably, e.g. due to clutter and partial occlusions. Such uncertainties cause erroneous detections when using a single-frame object detector only. Whilst temporal context can improve detections, the sparse and discrete nature of single-frame detectors (that use e.g. a non-maximum suppression to

R. Mester and M. Felsberg (Eds.): DAGM 2011, LNCS 6835, pp. 406–415, 2011.

sparsify detections) are unsuitable for this purpose, as too much information is discarded. In contrast, their dense and continuous raw detection-scores are more suitable for modeling the temporal context. In our work, we propose an efficient way to explore the spatial-temporal volume of dense detection-scores and extract tracklets from this volume. With an inhomogeneous Poisson process [8], we describe the tracklets as multi-frame observations without confidences originated from the corresponding tracks. Thus, based on the density of tracklets, the number of targets as well as their states can be estimated in a multi-frame tracking framework.

This paper presents three main contributions. Firstly, an efficient method to explore the spatial-temporal context of targets based on the continuous detection-scores is introduced in Section 3.1, Secondly, a new tracking framework based on multi-frame observations is proposed. In this framework, we use the density of tracklets rather than their detection-confidences as measurements for multi-frame tracking, to avoid the problem of instable detection-confidences due to occlusions and cluttered background. Finally, the experiments (Section 4) show a significant improvement of detection performance by employing tracklet detections. Our tracking algorithm outperforms the state-of-the-art method in crowded scenes.

2 Related Work

Significant progress in object detection recently has motivated research interest in detector-based tracking. Some approaches resolve the tracking problem by associating tracklets, i.e. short measurement sequences, using a global matching algorithm, e.g. [1,10]. These approaches are globally optimal in the sense of maximum-a-posteriori (MAP) probability. But they are often unsuitable for time-critical online applications. Moreover, detection-scores below the threshold-value are discarded, which could have been helpful information for linking tracklets. Wu and Nevatia [18] process online-tracking by associating detection responses of multiple confidence levels, which are generated by varying the thresholds. This algorithm is actually a compromise between efficiency (hard decision) and completeness (continuous scores).

Other algorithms make use of the intermediate output of detectors for tracking directly on the dense detection-score volume (TBD). In this case, detection decision, e.g. from thresholding, is delayed and tracking is performed on the raw measurements. It was first proposed for tracking weakly detected objects in radar applications where the SNR is low [15]. Recently, TBD has also been used for detector-based tracking. One category is online-boosting based tracking, e.g. [2]. They employ supervised or semi-supervised learning methods to distinguish the specific object from the background. However, drifting as a result from self-enforced wrong updates remains an issue. Another category is based on detectors trained offline [5], [14], [6]. In the work of Breitenstein et al. [5], they integrate the continuous detection-scores in a particle-filter framework. However, to avoid the exponential growth in the number of particles needed to represent the joint state space, they use independent particle sets for each target, which may

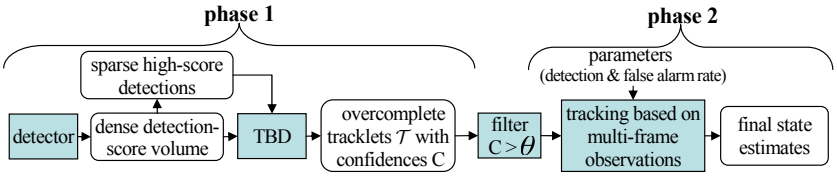


Fig. 1. The pipeline of our tracking system. TBD stands for tracking-by-detection.

have problems with interacting objects and occlusions. [14] generates a global spatial-temporal volume by combining detection-scores with ground-plane and background information. In this confidence volume, they apply a particle-filter to find the trajectories. Due to the lack of individual information in detection responses, such an optimization on responses over a long time is not robust against frequent interactions or occlusions between targets.

Building on the idea of tracklets, e.g. used in [1], we explore the local spatial-temporal volume of dense detection-scores in the form of tracklets, which are shown to improve detection significantly. As long tracks are not considered, it is more robust against occlusions. Based on the density of these tracklets, instead of the instable detection-confidences, a multi-frame tracking framework is proposed to estimate the target states.

3 Multi-target Tracking Based on Dense Detection-Scores

Fig. 1 depicts our tracking framework which consists of two phases. First, overcomplete tracklets are extracted by exploiting the local spatial-temporal context of detection-scores. Then, based on the density distribution of tracklets, we propose a Bayesian framework to estimate the target states jointly.

3.1 Generation of Tracklets

In many of the previous detector-based approaches, the high-score detections are used as input for tracking. However, these detections are sparse and unreliable in situations such as occlusions or complex backgrounds. For example, the detection-score often decreases when the object is partially occluded and thus discarded by thresholding. But in the temporal context, these responses are important cues for detecting the object completely and accurately over frames. On the other hand, some stochastically generated false alarms, which usually have high detection-scores for only one or two frames, can also be characterized by their large changes of the detection-scores over a short duration. In these situations, a frame-based threshold discards too much useful information. Thus, instead of using single-frame detections, tracklets $\mathcal{T}$ are extracted from the response volume as observations for tracking, where each tracklet $T \in \mathcal{T}$ is a sequence of target states $\{x_k, x_{k+1}, \cdots, x_{k+l-1}\}$ with a fixed length l.

We apply the detector on the input images $\mathcal{I} = \{\mathbf{I}_1, \mathbf{I}_2, \cdots, \mathbf{I}_k\}$ and obtain the detection responses $\mathcal{F} = \{\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_k\}$, where $\mathbf{F}_t$ are the responses at frame

t. Ideally, tracklets $\mathcal{T}$ could be determined from the response volume $\mathcal{F}$ jointly. However, such a global optimization on $P(\mathcal{F}, \mathcal{T})$ is computationally expensive, especially when the number of targets is high. As a compromise, we determine an overcomplete set of tracklets, where each tracklet $T \in \mathcal{T}$ is found separately in a local response volume. Intuitively, such local volumes should be chosen around high-score detections. More specifically, for each high-score detection d at frame t, a temporal window $[t, t+l-1]$ and a spatial neighborhood, depending on the assumed maximal velocity of a target (see (3)) and the size of the temporal window, are specified. In this spatial-temporal volume, we define the observation of a state x as a small spatial neighborhood around x at each frame s, denoted as $\mathbf{F}_{s|x}$. Let $\mathcal{F}_{t,d}$ be the set of the frame observations $\mathbf{F}_{s|x}$ in the local spatial-temporal volume, a tracklet is determined by maximizing the joint probability:

$$T^* = \arg\max_T P(\mathcal{F}_{t,d}, T) \; . \tag{1}$$

Using a hidden Markov model of first order, i.e. assuming state x_s is only dependent of the previous state x_{s-1}, the joint probability can be reformed as

$$P(\mathcal{F}_{t,d}, T) = P(\mathcal{F}_{t,d}|T)P(T) = \underbrace{\prod_{s=t}^{t+l-1} P(\mathbf{F}_{s|x}|x_s)}_{observation} \underbrace{\prod_{s=t+1}^{t+l-1} P(x_s|x_{s-1})}_{transition\ prob.} \underbrace{P(x_t)}_{init.\ prob.} \; . \tag{2}$$

Initial probability models the a-priori state distribution of objects. The scene knowledge can be modeled in this distribution. Without explicit information, we set $P(x_t)$ as a uniform distribution.

Transition probability is modeled by a gating function, which assumes a maximum velocity δ of targets:

$$P(x_s|x_{s-1}) = \begin{cases} 1/c & \text{if } x_s \in \Delta(\delta, x_{s-1}) \; , \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $\Delta(\delta, x_{s-1})$ is a spatial neighborhood of x_{s-1} defined by δ and $c = |\Delta(\delta, x_{s-1})|$ is the normalization factor.

Observation Model is approximated by the maximal response f_r from $\mathbf{F}_{s|x}$ and $P(f_r|x_s)$ can be learned by fitting the responses on a validation dataset.

Thus, optimal tracklets initialized by the high-score detections can be determined by using the Viterbi Algorithm. In this way, weak detections between high-score detections can be recovered and only a relatively low joint probability $P(\mathcal{F}_{t,d}, T)$ is assigned to 'isolated' false alarms. By setting a threshold θ on $P(\mathcal{F}_{t,d}, T)$, such false positive detections can be eliminated. As the optimization in a local volume is sensitive to initializations (starting frames and positions), overcomplete tracklets are generated, i.e. initialized from all high-score detections. In this way, our approach is robust against missing or inaccurate detections (see Fig. 3). Besides, dense tracklets would be generated in volumes with "continuous" high scores. Hence, the density of tracklets provides an alternative to detection-confidences as measures for the state-estimation.

3.2 Multi-target Tracking from Tracklets

The tracklets generated as described previously are an overcomplete set of track fragments. The remaining problem considered in this section is to estimate the number of targets and their tracks based on the observations of overlapping tracklets. To this end, we propose an approach similar to multiple hypothesis tracking (MHT) [7,4], which is a multi-target tracking with deferred decision. In our tracking model, tracklets $\mathcal{T}$ are considered as multi-frame observations and an inhomogeneous Poisson model [8] is used to describe the generation process of the set of tracklets from the tracks to be estimated. The main difference to conventional data-association based methods like JPDA [13,11] is that no explicit association has to be made. In this model, the density of tracklets, instead of their detection-confidences, are used as measurements, because of the discrete nature of the Poisson process. Besides, detection-confidences are instable due to e.g. occlusions, clutters, or bad illumination.

The advantage of multi-frame observations is that they provide clues for not only positions of targets at each frame, but also their temporal developments. Because tracklets are overlapping, it is easier to identify related tracklets than in the case of single-frame detections. Thus, the number of hypotheses that must be considered is reduced. This leads to a lower computational complexity.

Problem formulation: The multi-target state is $\mathcal{X}_t = \{X_t^1, \cdots, X_t^n\}$, where n is number of targets up to frame t and each track X_t^i is a sequence of target states. The observations are the tracklets $\mathcal{T}_t = \{T_t^1, \cdots, T_t^m\}$ from Section 3.1. The standard Bayesian formulation is applied to update the belief about the multi-target state:

$$P(\mathcal{X}_t|\mathcal{T}_t) \propto P(\mathcal{T}_t|\mathcal{X}_t)P(\mathcal{X}_t) \; . \tag{4}$$

In the following, the observation model and the prior model are detailed. For notational simplicity we omit the time index t.

Observation model: The inhomogeneous Poisson point process is used to model the likelihood function [8]. It assumes that the received observations $\mathcal{T}$ are generated by (conditionally independent) superposition of observations from $n+1$ sources, of which n are the targets $\mathcal{X}$ and the extra one is for the background clutter. Compared to the conventional likelihood model, complexity is significantly reduced as no explicit association between targets and observations is made and multiple observations originating from a target are allowed. Both the number of the observations and their spatial distribution are considered in this model. The joint likelihood of m observations $\mathcal{T} = \{T^1, \cdots, T^m\}$ for multi-target $\mathcal{X} = (X^1, X^2, \cdots, X^n)$ are [8]:

$$P(\mathcal{T}|\mathcal{X}) = \frac{e^{-\mu}}{m!} \prod_{j=1}^{m} \lambda(T^j|\mathcal{X}) = \frac{e^{-\mu}}{m!} \prod_{j=1}^{m} \sum_{i=0}^{n} \lambda_i(T^j|X^i) \; , \tag{5}$$

where $\mu = \sum_{i=0}^{n} \mu_i$ is the number of expected observations in the image area A, μ_i is the expected observations from target X^i, and $\lambda_i(p|X^i)$ describes the spatial density of observations in A with $\mu_i = \int_A \lambda_i(p|X)dp$. λ_0 models the observations

from clutter. We assume λ_0 uniformly distributed in A, i.e. $\lambda_0(p|X^0) = \rho$. This model was originally defined for single-frame observations. It can be easily extended for multi-frame observations:

$$P(\mathcal{T}|\mathcal{X}) = \underbrace{\frac{e^{-M}}{(m \cdot l)!}}_{\text{expected \#observations}} \underbrace{\prod_{j=1}^{m} \left(\rho^l + \sum_{i=1}^{n} \Lambda_i(T^j|X^i) \right)}_{\text{observation spatial density}} , \tag{6}$$

where l is the length of tracklets and $M = \sum_i M_i$ is the expected number of observations originating from targets $\{X^i\}_{i=1\cdots n}$. Let $A_{i,j}$ be the intersection frames of X^i and T^j. The multi-frame observation spatial density $\Lambda_i(T^j|X^i)$ is

$$\Lambda_i(T^j|X^i) = \rho^{l-|A_{i,j}|} \prod_{s \in A_{i,j}} M_i g(z_s|x_s) , \tag{7}$$

where $g(z_s|x_s)$ assumes a Gaussian distribution of observations z_s around the corresponding target x_s at each frame s. Substituting (7) into (6), we have

$$P(\mathcal{T}|\mathcal{X}) = c_1 e^{-M} \prod_{j=1}^{m} \left(1 + \sum_{i=1}^{n} \prod_{s \in A_{i,j}} c_2 g(z_s|x_s) \right) , \tag{8}$$

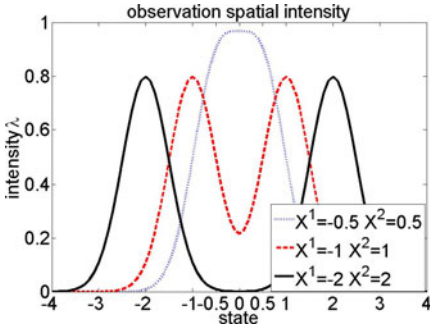
where c_1 is a constant independent of the tracks $\mathcal{X}$ and $c_2 = \frac{\mu_i}{\rho}$ models the signal-to-noise-ratio. In the implementation, the expected number of observations for each target is determined by $M_i = |X^i| \cdot l \cdot P_d$, where P_d is the detection rate.

However, for two very close targets, i.e. not only at one frame but several frames in our case, there is a problem of merged observations (see Fig. 2(a)). This leads to an ambiguity in the state estimation regarding the number of targets and the distribution for each target. In Fig. 2(b), two tracks and several tracklets (multi-frame observations) are visualized. The spatial intensities of tracklets $\sum_i \Lambda_i$ is indicated by the line width. While the overlap with a track increases the spatial intensity (e.g. T^1 and T^2, either a spatial deviation (T^3) or a temporal inconsistency (T^4) reduces the spatial intensity. It shows that tracklets provide strong clues in spatial as well as in temporal dimension for state estimation.

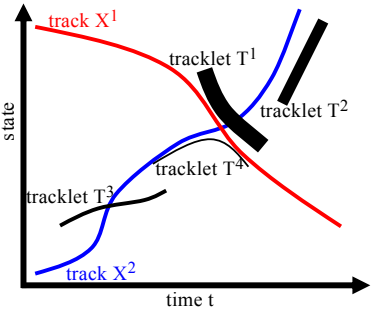
Prior probability: By assuming an independent and constant-velocity motion of targets, we penalize sudden changes in velocity:

$$P(\mathcal{X}) = \prod_{X \in \mathcal{X}} P(X) = \prod_{X \in \mathcal{X}} P(\|X^{''}\|_\infty) , \tag{9}$$

where $X^{''}$ the second-order motion-vector , i.e. accelerations at each frame, and $\|\bullet\|_\infty$ is the maximum norm. A Gaussian function $G(0, \sigma_m)$ is used to model the prior probability $P(\|X^{''}\|_\infty)$. The parameter σ_m is learned from some training sequences. Accelerations are normalized according to the target size, so that they are invariant to the 3D projection. Intuitively, $\|X^{''}\|_\infty$ increases as tracks get longer. Therefore, different σ_m are learned for different track lengths.



(a) Observation spatial intensity curves for two close (blue), moderately separated (red) and distant (black) targets (single frame). Observations are unresolved if targets are too close.



(b) Examples of tracks and tracklets. The spatial and temporal consistency of tracklets to tracks is important for the observation spatial intensities $\sum_i \Lambda_i$ (visualized by the line width of tracklets).

Fig. 2. Examples of observation spatial intensity for (a) single- and (b) multi-frame observations. We assume a Gaussian distribution of observations around the corresponding target.

State estimation: After specifying observation likelihood (8) and prior probability (9), multi-target states $\mathcal{X}$ can be estimated by maximizing the posterior distribution in (4). Let O be all frame-states of tracklet $\mathcal{T}$. A possible solution of $\mathcal{X}$ is a set of tracks $\{X^1, \cdots, X^n\}$, where each track X^i consists of a sequence of states from O and n is an unknown variable. Additionally, a maximum velocity of any target is assumed to reduce the solution space. For the reason of efficiency, the Markov chain Monte Carlo (MCMC) approach is employed to sample instead of enumerating all possible solutions. Similar to the multi-scan MCMCDA algorithm in [11], a variation of transition types are defined to initialize, terminate, split, merge, extend, reduce and switch tracks by sampling with the MCMC algorithm. As multi-frame observations are used, the convergence rate is fast. For example, about 2000 iterations are sufficient to track more than 20 targets in our experiments. Furthermore, a temporal window $[t-L+1, \cdots, t]$ of size L can be set, so that only the last L frames are revisited for the estimation.

4 Experiments

In this work, a cascade Adaboost classifier [17] trained on head-and-shoulder-patches is used, as the head-and-shoulder part of most people is visible in crowded scenes. For the evaluation, we use video sequences of crowded scenes from PETS2007 [16], MCTTR [9] and PETS2009 [12].

Detection Performance: First, we want to show the improvement of detection performance by using the proposed tracklets (the first phase of our algorithm, Section 3.1). In this experiment, we vary the length of tracklets. To compare

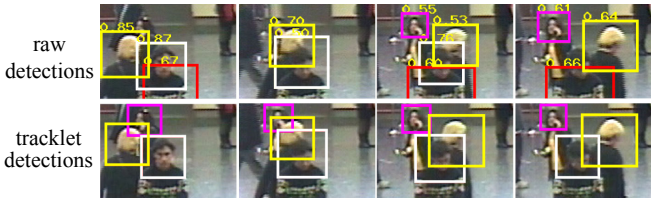
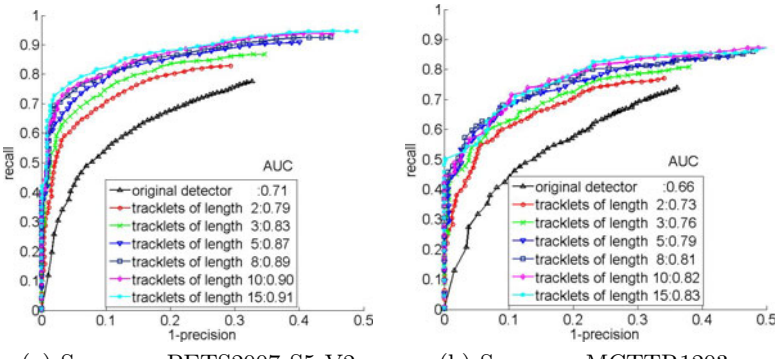


Fig. 3. Raw detections (first row) vs. tracklet detections (second row). The detection-scores of the objects (numbers above bounding boxes) are not stable, which leads to missing detections. These missing detections are recovered by our tracklet approach. Furthermore, false alarms are also reduced.

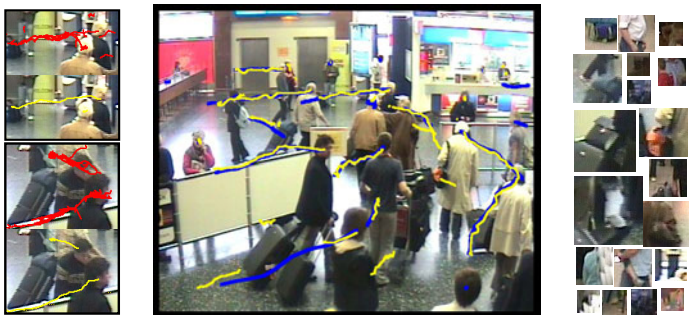


(a) Sequence PETS2007_S5_V2 (b) Sequence MCTTR1203a

Fig. 4. Comparison of detection-performance of tracklets and the original detector (black lines). Tracklets achieve much better results than the original detector. The improvement increases with the length of tracklets.

with the original detector, precision-recall curves are computed by adjusting the threshold value for the joint probability $P(\mathcal{F}_{t,d}, T)$. In Fig. 4, recall rates are significantly improved by tracklets which exploit the temporal context of strong detections. Further, many false positives of the original detector, even with high detection-scores (left-side of the curve), are eliminated. It proves the effectiveness of tracklets by removing the sporadic false alarms with high detection-scores. The performance increases also with the length of tracklets. However, the improvements saturate at length 10.

Tracking Performance: The proposed tracking algorithm is based on tracklets detections. Fig. 5 shows some tracking results from our complete two-phase approach proposed in Section 3. By modeling false alarms explicitly in the likelihood model and introducing the motion model, our approach is robust against some outliers in the generated tracklets (Fig 5(a)). However, some false alarms still remain (Fig. 5(c)). Most of them do not change appearance much over time, hence relatively constant detection-responses are obtained. In this situation, the proposed approach can not distinguish them from the correct target, as the detector is the only source of information for tracking. Combination of different detectors could probably alleviate this problem.



(a) (b) (c)

Fig. 5. (a) Tracking output (yellow lines) from tracklets (red lines). The track estimates are robust against the outliers in tracklets. (b) Tracking output (yellow lines) are compared to the ground truth (blue lines). Targets far from the camera (too small) are not tracked well. (c) Examples of false alarms by tracking. Most of them have relatively constant detection-responses, i.e. with stable appearance over time.

Table 1. Quantitative evaluation. Compared to the results in [6] on PETS2009, our algorithm has much better precision (MOTP) and similar accuracy (MOTA).

Seq.	HD RailwA 	HD RailwB 	PETS2007 S5-V2 	MCTTR 1203a 	PETS2009 S2-L2 	PETS2009 S2-L3 
MOTP	83.8%	82.7%	81%	78.8%	79.1% (51.3% [6])	80.1% (52.1% [6])
MOTA	83.9%	72.4%	72.6%	60.7%	55.1% (50.0% [6])	61.0% (67.5% [6])

The CLEAR MOP metrics [3] are used to evaluate the tracking performance quantitatively. The precision score MOTP (intersection over union of bounding boxes) and the accuracy score MOTA (composed of false negative rate, false positive rate and number of ID switches) are computed. The results are shown in Table 1. We also compare our method with the results reported in [6] for PETS2009. Our algorithm achieves a much higher precision score (MOTP). It benefits from our head-shoulder model (suffering less occlusion problems in such crowded scenes) and the exploitation of the dense spatial-temporal volume. The accuracy (MOTA) of our algorithm is similar to that of [6]. Our tracking system is implemented in C++. The runtime depends on the number of targets and tracklets. For the sequence MCTTR1203a with about 20 targets, the processing time of tracking is about 0.8 second/frame. Further optimization is possible.

Summary: The proposed tracklet approach improves the detection performance significantly. Based on that, our algorithm provides robust tracking in challenging crowded sequences and outperforms a state-of-the-art method.

5 Conclusions

In this paper, we propose a novel tracking framework based on the dense output of the detector. From local spatial-temporal volumes of dense detection-scores, tracklets are extracted to improve the detection performance. Instead of using detection-confidences directly, which are usually instable due to occlusion and clutter, overcomplete tracklets are generated and their density is considered as measurements for tracking. By modeling the tracklets and their density with an inhomogeneous Poisson process, target states are estimated efficiently in a Bayesian tracking framework. Compared to the state-of-the-art method, our algorithm achieves better tracking results in crowded scenes.

References

1. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: Proc. CVPR (2008)
2. Avidan, S.: Ensemble tracking. In: Proc. CVPR (2005)
3. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. Journal Image Video Process 2008, 1–10 (2008)
4. Blackman, S.: Multiple hypothesis tracking for multiple target tracking. IEEE Trans. on Aerospace and Electronic Systems 19(1), 5–18 (2004)
5. Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Robust tracking-by-detection using a detector confidence particle filter. In: ICCV (2009)
6. Breitenstein, M. D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Online multi-person tracking-by-detection from a single, uncalibrated camera. IEEE Trans. on Pattern Analysis and Machine Intelligence PP(99), 1 (2010)
7. Cox, I.J.: A review of statistical data association techniques for motion correspondence. International Journal of Computer Vision 10(1), 53–66 (1993)
8. Gilholm, K., Godsill, S., Maskell, S., Salmond, D.: Poisson models for extended target and group tracking. In: Proc. SPIE (2005)
9. Home Office: Multiple camera tracking scenario data, `http://www.homeoffice.gov.uk/science-research/hosdb/`
10. 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)
11. Oh, S., Russell, S., Sastry, S.: Markov chain monte carlo data association for multi-target tracking. IEEE Trans. on Automatic Control 54(3), 481–497 (2009)
12. PETS workshop: PETS (2009), `http://www.cvg.rdg.ac.uk/PETS2009/`
13. Roecker, J.: A class of near optimal jpda algorithms. IEEE Trans. Aerospace and Electronic Systems 30, 504–510 (1994)
14. Stalder, S., Grabner, H., Van Gool, L.: Cascaded confidence filtering for improved tracking-by-detection. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 369–382. Springer, Heidelberg (2010)
15. Tonissen, S., Evans, R.: Peformance of dynamic programming techniques for track-before-detect. IEEE Trans. on Aerospace and Electronic Systems 32(4), 1440–1451 (1996)
16. UK EPSRC REASON Project: PETS (2007), `http://pets2007.net/`
17. Viola, P., Jones, M.: Robust real-time object detection. International Journal of Computer Vision 57(2), 137–154 (2002)
18. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. International Journal of Computer Vision 75(2), 247–266 (2007)