Exploring Bounding Box Context for Multi-Object Tracker Fusion

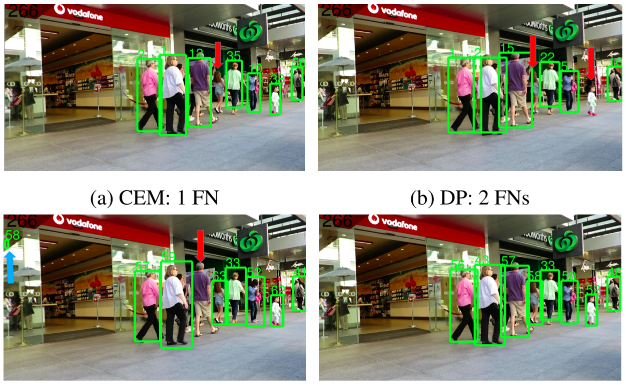
Stefan Breuers[1] Shishan Yang[2] Markus Mathias[1] Bastian Leibe[1]

[1]Visual Computing Institute
RWTH Aachen University
{breuers,mathias,leibe}@vision.rwth-aachen.de

[2]Data Fusion Group
Georg-August-Universität Göttingen
shishan.yang@cs.uni-goettingen.de

Abstract

Many multi-object-tracking (MOT) techniques have been developed over the past years. The most successful ones are based on the classical tracking-by-detection approach. The different methods rely on different kinds of data association, use motion and appearance models, or add optimization terms for occlusion and exclusion. Still, errors occur for all those methods and a consistent evaluation has just started. In this paper we analyze three current state-of-the-art MOT trackers and show that there is still room for improvement. To that end, we train a classifier on the trackers' output bounding boxes in order to prune false positives. Furthermore, the different approaches have different strengths resulting in a reduced false negative rate when combined. We perform an extensive evaluation over ten common evaluation sequences and consistently show improved performances by exploiting the strengths and reducing the weaknesses of current methods.



(a) CEM: 1 FN

(b) DP: 2 FNs

(c) TBD: 1 FN, 1 FP

(d) SVM-Union: no errors

Figure 1: Qualitative example of different mistakes of different tracking systems. FNs are marked with a red arrow, FPs with a blue one, respectively. Note that the combined result has correctly discarded the FPs and filled in all FNs by combining the results.

1. Introduction

Over the past decade, Multi-object-tracking (MOT) has been an area of active research [22, 24, 25, 29, 20, 19, 15, 10, 9, 30, 13, 28]. Many new approaches are proposed every year with applications in robotics, video analysis, and autonomous vehicles.

The most common approaches make use of object detectors to identify, *e.g*., pedestrians in images. Given potential object positions in each image/frame of a video sequence, tracking-by-detection approaches perform data association to group the detections into tracks and, at the same time, try to identify failures, *i.e*., missing detections and false alarms. Often, the tracking result includes assumptions about an underlying motion or appearance model and an explicit handling of target exclusion or occlusion.

There is a considerable interest in determining how well MOT techniques perform [14] and where mistakes are made [21]. The main question here is how far are we from saturation and is there still potential for improvement? In this paper, we show that MOT performance has not yet saturated.

We analyze several key aspects which often result in tracking errors for multi-person tracking scenarios. Detectors often fail for very small objects or objects that are large but close to image borders. In crowded scenes, not only do the detectors encounter problems, *e.g*., due to overlapping and occluding pedestrians, but also tracking gets more difficult, *i.e*., data association may fail. This is reflected in the bounding boxes of the tracking result. Even when ignoring the internal states of the various tracking systems, we are able to improve tracking performance while only looking at the output bounding boxes. Two approaches are used to improve the performance. First, for each method, we train a false positive classifier based on few additional features. This already boosts the trackers' precision. Then, we exploit that different trackers can compensate for each others' mistakes. By combining tracking results, false negatives (FNs) are reduced, resulting in a better recall.

Our main contributions are: (1) We show that current MOT approaches can be improved on the basis of their predicted 2D bounding boxes. (2) We present a new FP-classifier to improve tracking scores of individual trackers. (3) We show that combining trackers helps to compensate for individual mistakes, setting a new state-of-the-art performance.

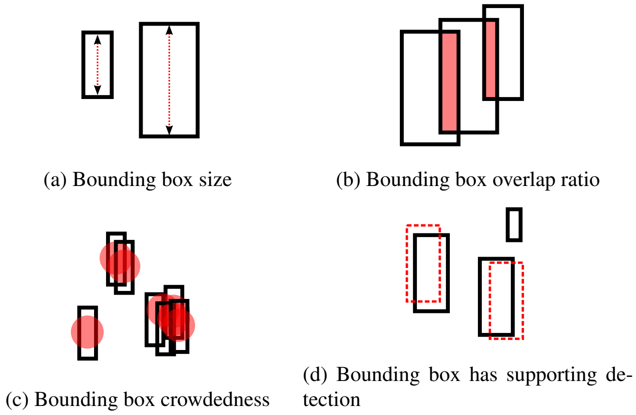
The paper is structured as follows. We first review related work (Section 2). Section 3 introduces our false positive classifier. In Section 4 we present our approach to combine different trackers. Experiments and evaluation are given in Section 5.

2. Related Work

The problem of MOT branches into many problem formulations, each with many methods, so we want to put the focus on visual pedestrian tracking-by-detection.

The most common approach for tracking to deal with the detector input is the Bayesian probability formulation. These methods formulate tracking as inference, trying to solve it via maximum a posteriori (MAP) estimation. Often an Extended Kalman Filter (EKF) [22] or a Particle Filter [24] is utilized. The weakness of the Particle Filter is the computational complexity, while the result of the (E)KF depends on the underlying motion model. Another branch are graph-based approaches, like [25], where the detections form a network and a tracking result is found via maximum flow. Further methods, like [29] use a CRF of small tracklets. The result of graph-based approaches sticks to detection, but using motion or occlusion information is often problematic. Here other approaches work better, *e.g.*, the energy-minimization of Milan *et al.* [20, 19], which is closely related to MCMCDA by [23]. Other methods reach from Quadratic Pseudo-Boolean Optimization [15], the Hungarian algorithm [10, 9, 30], using person identity [13] or motion agreement [28].

Not only the number of methods makes evaluation difficult, but also other problems, like label noise [21] and all the possible errors that can occur. While, like in detection, false positives (FPs), and false negatives (FNs) can be used to compute precision or recall, respectively, also ID switches (IDS) or fragmentations (FRA) are of interest. With the widely used CLEAR metrics [4] a more compact evaluation is possible. Here Multi Object Tracking Accuracy (MOTA) represents the ratio of errors (FPs, FNs, IDS) to ground-truth targets and Multi Object Tracking Precision (MOTP) the average bounding box accuracy (average 2D IoU) of matched pairs. Another tracking score proposed by Li *et al.* [16] captures the mostly tracked (MT: more than 80%), mostly lost (ML: less than 20%) and partially tracked (PT) ground truth targets. Still, it is important to keep all the other factors (annotations, detections, evaluation script,...) consistent [21]. The work on a consistent MOT evaluation has just started [14].


(a) Bounding box size

(b) Bounding box overlap ratio

(c) Bounding box crowdedness

(d) Bounding box has supporting detection

Figure 2: Illustration of the used 2D bounding box context features used to train the FP-classifier. Solid boxes represent tracker output.

Combining different tracking results has mainly been done for single object tracking. For single object or template based tracking already some fusion approaches exist, reaching from early feature fusion [31], to late tracker fusion, *e.g.*, using HMM [27] or SVMs [26], combining the results of different trackers and showing an overall improvement. [18] and [3] present single-object tracker fusion methods on bounding box level, where [18] compares averaging, median and majority voting, and [3] uses dynamic programming to maximize an attraction energy function finding a continuous trajectory. Both methods only take into account the position and size of bounding boxes of different single-object trackers, where in our case overlap, crowdedness and detection support are taken as additional features in the MOT domain. Yet, there is hardly any work carried out on fusing MOT results. Nevertheless, we do not claim to give a sophisticated MOT fusion approach, but rather show that in general the results of different tracking system are in some ways orthogonal and can be combined to compensate for each others mistakes, when keeping an eye on the balance of errors. There is still room for improvement, based on the used scene information, *e.g.*, bounding box context information to improve the behavior of different MOT trackers.

3. FP classification with bounding box context

In this section we propose to identify tracking errors based on the proposed output bounding boxes. Through the observation that one tracker tends to struggle in similar conditions (*e.g.* for small pedestrians, or in crowded scenes) our goal is to learn to identify such situations. Ideally, different trackers have different strengths, such that by removing individual weaknesses and then combining the trackers will result in an improved performance. Therefore the false positive classifier is trained for each tracker individually, to exploit the individual tracker behaviors originating from the motion model, occlusion terms, etc. and the failures that may arise because of those.

Inspired by [17] we have chosen the following features to capture bounding box context information:

Bounding box size. Tracking pedestrians far away from the camera tends to be more difficult. In 3D tracking the projected world position of a 2D bounding box belonging to a person further away differs by several meters, if the bounding box is only shifted by one pixel. In 2D, a smaller bounding boxes contain fewer information, influencing, *e.g.*, the motion, or appearance model. For some trackers, also too large bounding boxes may be a cause of error. We use the bounding box height to approximate the camera-pedestrian distance.

Bounding box overlap. As soon as an object gets partly or fully occluded by other targets, keeping track of this object gets harder. Unlike detectors, trackers are supposed to localize those objects but tend to make mistakes. We use the overlap-area-ratio to quantify if a tracked bounding box might be occluded by other bounding boxes, disregarding the exact role of the occlusion (occluder or being occluded).

Bounding box crowdedness. Tracking and identifying individual pedestrians in crowded spaces is challenging. Based on the detector output and the tracking history, the tracker needs to quantify the number of objects in the scene and also estimate their movement. Especially in crowded scenes, the individual movements might drastically change, *e.g.*, to avoid collisions and might not be well predicted by the motion model. Although bounding box overlap and crowdedness are correlated, experiments show that using both features is beneficial. To approximate the crowdedness a Gaussian window is centered on each tracked bounding box, with the doubled width as variance. The sum of those Gaussian weights at a center point of a bounding box serves as an indicator for the crowdedness.

Bounding box has detection support. This binary feature captures if one by the tracker proposed bounding box is supported by a detection or not. We consider a detection to support a 2D tracking bounding box, if the IoU, *i.e.*, the Jaccard index, exceeds 50%.

Based on these four features, we train a support vector machine (SVM) with a RBF-kernel individually for each tracker in order to classify the predicted bounding boxes as correct (TP) or not (FP).

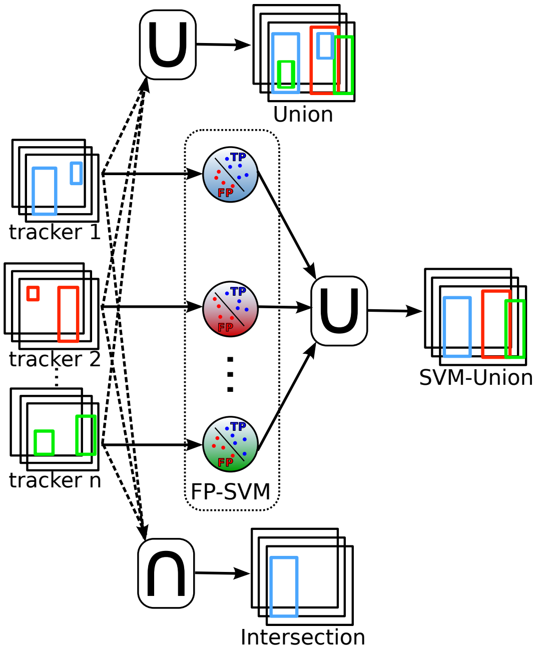


Figure 3: Overview of the three different combination approaches.

4. Tracker combination

Different tracking systems often make different mistakes. For a qualitative example of such different errors see Fig. 1. The reason for this can be explained by the usage of different data association, motion and appearance models, or if a method explicitly handles occlusions or not.

In the last section we analyzed bounding box context to exclude FPs from individual trackers. We will now combine different trackers with the goal of compensating each others mistakes. We will present two baselines fusion approaches together with our proposed tracker combination using the false positive classifier from Section 3. The overview of the different fusion approaches is illustrated in Fig. 3.

4.1. Union and Intersection

First we introduce two naive approaches to combine the different trackers: bounding box Union and Intersection.

For the Union baseline we collect all the bounding boxes of the different trackers. Whenever two or more bounding boxes overlap, they are likely to describe the same ground truth object. In that case, we construct one new bounding box from all overlapping bounding boxes by averaging over position and size (width and height). Which tracking bounding boxes belong together is decided in a greedy data association step based on the Jaccard index (IoU>50%). Regarding the IDs of the combined targets a consistent ID is kept as long as one of the different trackers had no ID switch. The target IDs of the different trackers are stored in a simple map. As soon as all trackers change their ID regarding to this map, a new ID is assigned to the combined result. Intuitively, this Union of the different trackers should lead to a reduction of false negatives, as missed pedestrians from one tracker might be found by another. On the other hand all the errors add up. The focus of this baseline clearly lies on decreasing the FNs (better recall) but has the drawback of increasing the FPs (worse precision).

The Intersection baseline follows the opposite strategy by only keeping bounding boxes when all of the trackers agree on them. Bounding box averaging and ID-handling is carried out in the same way as for Union. Intersecting means that the number of false positives (better precision) will be decreased while missing out in recall.

4.2. SVM-Union

Union or Intersection both share the problem that one error (FP or FN, respectively) is decreased, while the other one increases. The so called MOTA score combines the rates of FPs, FNs and ID switches and therefore decreases for the proposed baselines. We consider this score crucial to assess the combined trackers performance.

Our proposed method takes into account both: the combined strength of all trackers to decrease FNs and also tries to avoid accumulating FPs.

Algorithm 1 SVM-Union

```
Input: trackers $T_{1...K}$ with result $bb^{T_k} = \{bb^{T_k}\}_{i,t}$
(box $i$ at frame $t$) and trained FP-SVM$^{T_k}$
Output: Combined bounding boxes $bb^*$

// Apply FP-SVM to prune individual results
 1: for each method $T_k$ do
 2:     $bb^{T_k} \leftarrow$ FP-SVM$^{T_k}(bb^{T_k})$

// Build Union on individual results
 3: $bb^* \leftarrow \emptyset$
 4: for each frame $t$ do
 5:     for each method $T_k$ do
 6:         for each $bb_{i,t}^{T_k} \in bb^{T_k}$ do
 7:             $bb' \leftarrow \{bb_{i,t}^{T_k}\}$
 8:             $bb^{T_k} \leftarrow bb^{T_k} - bb_{i,t}^{T_k}$
 9:             for each other method $T_l, (l \neq k)$ do
10:                 $j = \arg\max_z \text{IoU}(bb_{i,t}^{T_k}, bb_{z,t}^{T_l})$
11:                 if $\text{IoU}(bb_{i,t}^{T_k}, bb_{j,t}^{T_l}) > 0.5$ then
12:                     $bb' \leftarrow bb' \cup bb_{j,t}^{T_l}$
13:                     $bb^{T_l} \leftarrow bb^{T_l} - bb_{j,t}^{T_l}$
14:             assign_ID($bb'$)
15:             $bb^* \leftarrow bb^* \cup$ average_position($bb'$)
```

The main steps of the algorithm are shown in Alg. 1. First the individual tracking results are filtered using the learned classifier (Section 3). Then the Union operation is applied, combining the strengths of the different approaches. The IDs (line 14) are assigned with a simple map, as described in Section 4.1. Position and size of a combined box (line 15) are averaged. As will be shown in the results each tracker has some confident true positives, which are not removed by the SVM classifier but at the same time are not shared by the other trackers. Through such configuration we are able to improve the overall tracking result.

5. Experiments

For our experiments, we have chosen a representative set of MOT tracking methods, also based on code availability.

CEM. The continuous energy minimization by Milan *et al.* [20] designs an energy function and minimizes it by using a jump framework, closely related to MCMCDA. The energy is composed of several terms, incorporating observation support, constant velocity dynamics, spatial exclusion and persistent existence of targets, as well as their number and the length of their tracks [19]. This method leads to long consistent tracks, with lowest number of ID switches (IDS) compared to the other methods.

DP. The graph-based approach by Pirsiavash *et al.* [25] minimizes the network-flow via dynamic programming. As tracking is performed strictly on the detections without additional occlusion handling, DP gives short but reliable tracks, leading to a good precision (low number of FPs), but suffers from more misses than other methods (number of FNs).

TBD. The tracking-by-detection framework of Geiger *et al.* [10, 9, 30] is part of a complex tracking system in urban scenes. It relies on the well established extended Kalman Filter (EKF) for predicting the targets' positions in case of a missing detection. The Hungarian algorithm is used to both solve bipartite matching, associating detections from one frame to another and connecting tracklets employing an occlusion sensitive appearance model to bridge gaps. TBD is able to track many targets resulting in good recall with a lower precision than the other methods.

Note that all trackers are used as provided by the authors. No parameter tuning has been done, especially not to optimize the trackers' output for our classification or combination approach.

To ensure a consistent setup we use the DPM detections of Felzenszwalb *et al.* [12, 7] and the evaluation script and annotations from the MOTchallenge2015 [14] throughout all the experiments. The methods are compared by using common measures to assess the tracking quality, already mentioned in Section 2: CLEAR metrics of [4] (MOTA, MOTP), Recall, Precision, number of False Positives (FPs), number of False Negatives (FNs), Number of ID switches

	MOTA	MOTP	Recall	Precision	FP	FN	IDS	MT	PT	ML	FRA
CEM	25.73	71.29	35.70	83.84	244.9	2496.2	41.4	1.9	12.6	27.4	38.6
CEM+SVM	**27.74**	71.56	33.85	90.30	142.5	2581.7	42.8	1.3	12.2	28.4	53.4
DP	26.44	70.99	31.94	92.72	103.9	2665.6	81.8	0.8	13.3	27.4	93.1
DP+SVM	26.39	71.00	31.52	93.80	86.8	2695.4	79.9	0.7	13.1	27.7	93.0
TBD	36.52	71.31	47.01	85.65	288.4	2117.1	79.2	4.1	18.2	19.6	100.7
TBD+SVM	**37.35**	71.32	45.83	88.17	246.5	2176.5	75.7	4	17.7	20.2	101.4

Table 1: Comparison of single tracking results and single tracker+FP-SVM-classifier. Averaged tracking scores over all scenes. Only improvements are highlighted in red to keep the presentation clear.

	-% FP	-% TP	+% FN	MOTA better	MOTA worse	MOTA equal
CEM+SVM	41.81	7.43	3.43	**8**/10	2/10	0/10
DP+SVM	16.46	3.04	1.12	2/10	1/10	7/10
TBD+SVM	14.53	3.88	2.81	**6**/10	3/10	1/10
method avg.	24.27	4.78	2.45	16/30	6/30	8/30

Table 2: Evaluation of the SVM-classifier in respect to the relative decrease of false positives (FP), as well as the decrease of true positives (TP) and the involving increase of false negatives (FN). The resulting development in the MOTA score is listed via the number of sequences.

	MOTA	MOTP	Recall	Precision	FP	FN	IDS	MT	PT	ML	FRA
CEM	25.73	71.29	35.70	83.84	244.9	2496.2	41.4	1.9	12.6	27.4	**38.6**
DP	26.44	70.99	31.94	92.72	103.9	2665.6	81.8	0.8	13.3	27.4	93.1
TBD	36.52	71.31	47.01	85.65	288.4	2117.1	79.2	4.1	18.2	19.6	100.7
Union	37.2	71.6	**50.10**	84.10	336	**2028**	76	**6**	17	**19**	85.9
Intersection	23.2	**72.2**	25.60	**93.30**	**80**	2818	**15**	1	10.3	31	69.9
SVM-Union	**39.12**	71.66	48.88	87.70	257.5	2095.5	71.7	5.5	16.8	19.6	93.6

Table 3: Comparison of all single tracking results and the combinations. Averaged tracking scores over all scenes. Best scores are highlighted in bold red. Second best scores in red, to also show improvements without the obvious baselines.

(IDS), number of mostly tracked ($\geq$ 80%) pedestrians (MT), number of mostly lost ($\leq$ 20%) pedestrians (ML), partially tracked (>20% and <80%) pedestrians (PT), and the fragmentations (FRA).

We tested our methods on a large set of publicly available sequences of different data sets (ETH [6, 5], ADL [14], KITTI [11], TUD [2, 1], PETS [8], Venice [14]), covering the MOTChallenge2015 benchmark [14].

The training of the classifier (Section 3) is done via cross-validation, *i.e.*, in a leave-one-out fashion, to maximize the training information, while we test on the leftover sequence.

5.1. False Positive Classification

We first analyze the quality of the false positive classifier (FP-SVM) individually for each tracker and show its impact on the tracking scores. Table 1 summarizes the results, showing the individual methods alone and in combination with the FP-SVM. All scores are averaged over the ten sequences. The scores of the individual sequences can be found in the supplementary material. As can be seen on the left half of Table 2 the number of false positives are reduced by 24.3%, while removing only 4.8% of the good hypotheses (increasing the number false negatives by only 2.5%) averaged over all methods. This leads to a consistent improvement in precision and MOTP. It also shows that our false positive classification is effective.

On a sequence level, the right part of Table 2 shows the absolute number of sequences on which the FP-SVM improved results based on MOTA score, how often it got worse or stayed the same. For CEM and TBD the FP-SVM improved MOTA in most of the sequences. On the other hand DP does not severely change. This is good news, as Table 1 shows that DP has already a very good precision and the FP-SVM learns not to throw away good bounding boxes. This is also consistent with our observation that DP outputs short but precise tracks. For reference, the highest improvement of the MOTA score on a single sequence is +6.1 percentage points for TBD on KITTI-17, +9.9 for CEM and +1.2 for DP on KITTI-13.

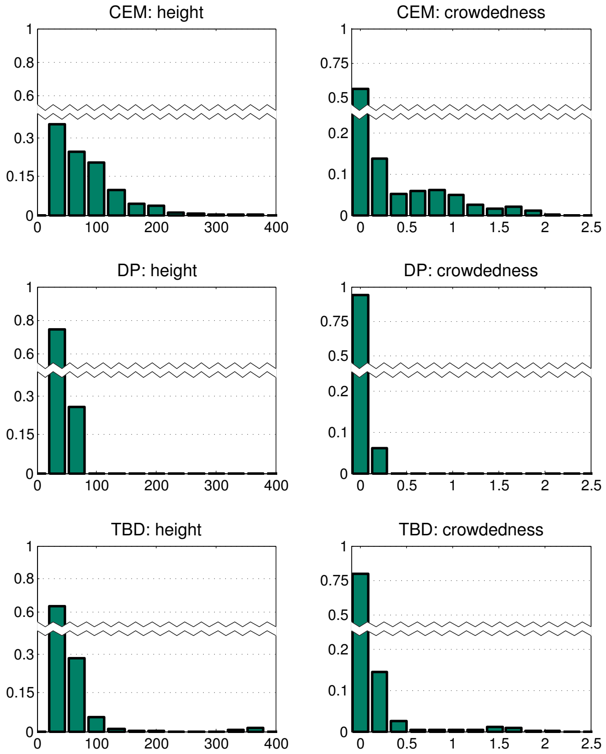


Figure 4: Normalized histograms of boxes that are classified as FP by the SVM classifier. The projection on the two features height (left) and crowdedness (right) are shown for each method, averaged over all sequences.

Fig. 4 gives some insight to the different classifiers' behavior, showing the distribution of boxes classified as FP on two projections in the feature space, height and crowdedness. The other two features, occlusion and detection support can be found in the supplementary material. We can see that the SVM for DP and TBD mainly discards small boxes in clear areas. While for TBD still some large boxes and boxes in crowded areas are discarded, the classifier learned for CEM captures a broader distribution of discarded boxes in terms of both heights and levels of crowdedness. Note that this projection onto one feature is simplistic but gives a rough idea of the learned failure cases of the individual tracking methods.

5.2. Tracker Combination

We proposed several combinations of trackers, namely Union, Intersection, and SVM-Union. Tab. 3 compares the combined methods to the original individual ones. We again show the averaged results over all sequences, results on individual sequences are found in the supplementary material.

The results of our baselines look somewhat expected. Intersection has the best precision, but the worst recall; Union the best recall but misses precision. The result also shows that our proposed SVM-Union combines the benefits of both of the (rather extreme) baselines. This can be mainly observed on the combined tracking scores MOTA (but also MOTP), for which our SVM-Union outperforms all of the individual methods. The best individual method (TBD) by 2.6 percentage points and CEM by 13.39 percentage points. On the sequence level, SVM-Union improves the MOTA score on 8 out of the 10 sequences. Fig. 5 gives an overview of the development of the MOTA score for the individual sequences. The comparison shows each original method, its SVM pruned classification result and the final fusion result, showing an overall improvement.

6. Conclusion

In this paper we have shown that current state-of-the-art MOT trackers can still be improved by only using the high level information of bounding box context. We have designed a FP-classifier which is (despite of its simplicity) able to improve tracking scores of the core trackers on a large set of commonly used sequences. In addition we analyzed the possibilities to combine several MOT results. Our final method SVM-Union is able to prune errors of the individual methods while combining their strength resulting in an improved tracking performance.

The proposed late fusion of the trackers shows that there is still room for improvement and that there are some orthogonal strengths and weaknesses of the individual methods which should be exploited in a unified tracking approach in future work.

Acknowledgement

This work was funded, in parts, by EU project SPENCER (FP7-ICT-600877) and ERC Starting Grant project CV-SUPER (ERC-2012-StG-307432).

References

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR*, 2008.

[2] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010.

[3] C. Bailer, A. Pagani, and D. Stricker. A superior tracking approach: Building a strong tracker through fusion. In *ECCV*. 2014.

[4] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008.

[5] A. Ess, B. Leibe, and L. V. Gool. Depth and appearance for mobile scene analysis. In *ICCV*, 2007.

[6] A. Ess, B. Leibe, K. Schindler, , and L. van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, June 2008.

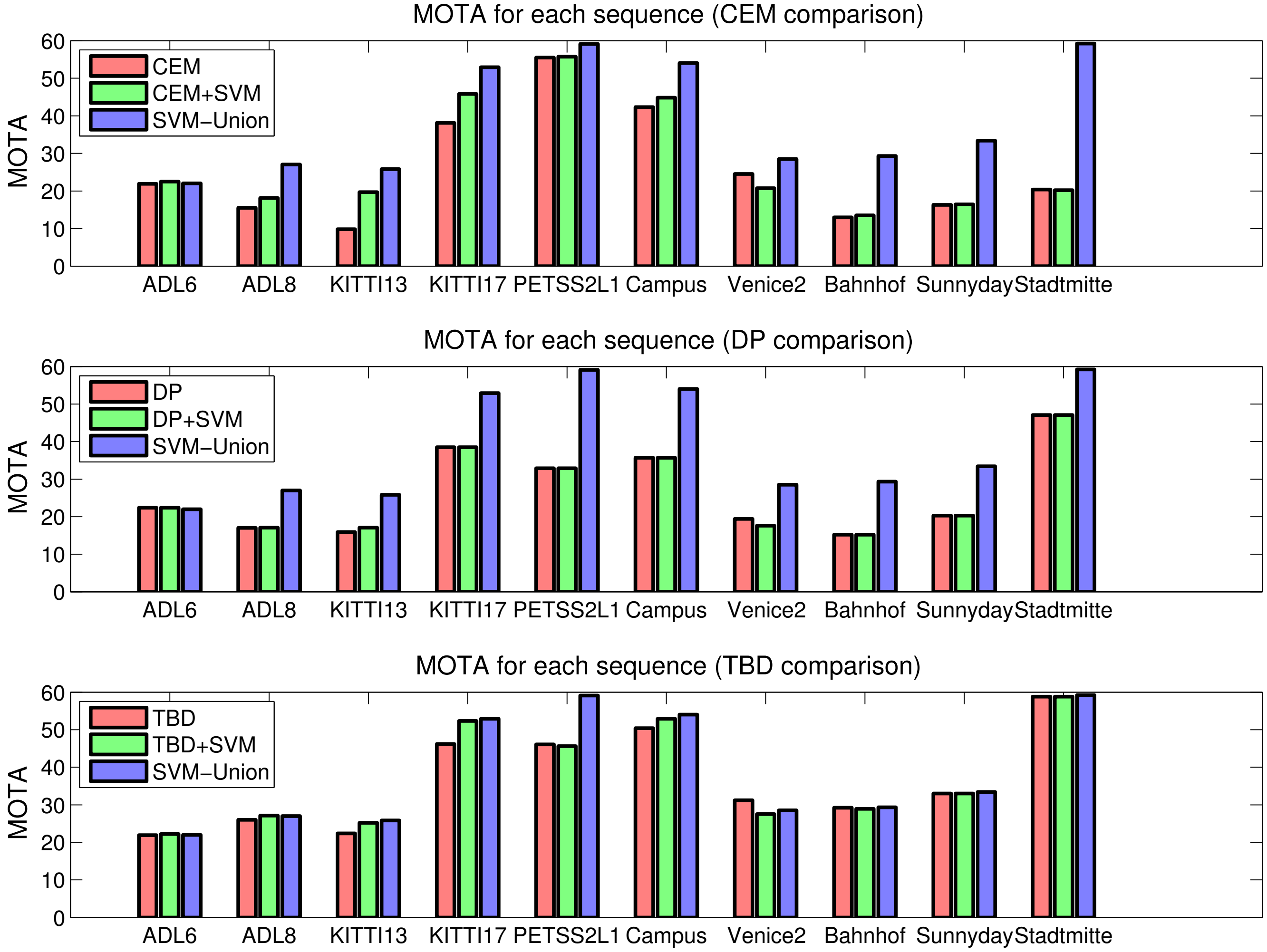


Figure 5: MOTA scores for each individual sequence. Comparison between single tracking result (red), with FP-SVM (green) and proposed SVM-Union (blue).

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), 2010.

[8] J. Ferryman and A. Shahrokni. Pets2009: Dataset and challenge. In *PETS*, 2009.

[9] A. Geiger. *Probabilistic Models for 3D Urban Scene Understanding from Movable Platforms*. PhD thesis, KIT, 2013.

[10] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3d traffic scene understanding from movable platforms. *PAMI*, 2014.

[11] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

[12] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/.

[13] C. Kuo and R. Nevatia. How does person identity recognition help multi-person tracking? In *CVPR*, 2011.

[14] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *CoRR*, 2015.

[15] B. Leibe, K. Schindler, N. Cornelis, and L. J. V. Gool. Coupled object detection and tracking from static cameras and moving vehicles. *PAMI*, 30(10), 2008.

[16] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *CVPR*, 2009.

[17] W. Luo, X. Zhao, and T. Kim. Multiple object tracking: A review. *CoRR*, abs/1409.7618, 2014.

[18] R. Martín and J. M. Martínez. Evaluation of bounding box level fusion of single target video object trackers. In *Hybrid Artificial Intelligence Systems*. 2014.

[19] A. Milan, S. Roth, and K. Schindler. An analytical formulation of global occlusion reasoning for multi-target tracking. In *Proc. of the 11th International IEEE Workshop on Visual Surveillance*, 2011.

[20] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *PAMI*, 36(1), 2014.

[21] A. Milan, K. Schindler, and S. Roth. Challenges of ground truth evaluation of multi-target tracking. In *CVPR Workshop*, 2013.

[22] D. Mitzel and B. Leibe. Real-time multi-person tracking with detector assisted structure propagation. In *ICCV Workshops*, 2011.

[23] S. Oh, S. Russell, and S. Sastry. Markov chain monte carlo data association for multi-target tracking. *IEEE Transactions on Automatic Control*, 54(3), 2009.

[24] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV*, 2004.

[25] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011.

[26] F. Tang, S. Brennan, Q. Zhao, and H. Tao. Co-tracking using semi-supervised support vector machines. In *ICCV*, 2007.

[27] T. Vojir, J. Matas, and J. Noskova. Online adaptive hidden markov model for multi-tracker fusion. *CoRR*, 2015.

[28] Z. Wu, J. Zhang, and M. Betke. Online motion agreement tracking. In *BMVC*, 2013.

[29] B. Yang and R. Nevatia. An online learned CRF model for multi-target tracking. In *CVPR*, 2012.

[30] H. Zhang, A. Geiger, and R. Urtasun. Understanding high-level semantics by modeling traffic patterns. In *ICCV*, 2013.

[31] Y. Zhou, C. Rao, Q. Lu, X. Bai, and W. Liu. Multiple feature fusion for object tracking. In *Intelligent Science and Intelligent Data Engineering*. 2012.