Introducing Target Profiling for Context-Aware Tracking

A. Dimou, A. Axenopoulos and P. Daras

Information Technologies Institute, Centre of Research and Technology - Hellas,
6th km Charilaou - Thermi, 57001, Thessaloniki, Greece, {dimou, axenop, daras}@iti.gr

Keywords: Tracker, context-aware, fusion, scene modelling.

Abstract

In this paper, an automated methodology that builds a profile for each pedestrian tracked based on its appearance, its occlusion status and the semantic information related to its position, is presented. The extracted profiles are utilized to perform context-aware tracking in multi-target tracking scenarios. A novel fusion scheme that combines the output of multiple trackers, exploiting context-related information cues is proposed. A set of decision rules is created that implicitly integrates occlusion reasoning capabilities in multi-target scenarios. Key aspects of the fusion process presented are (a) a common, context-aware methodology to assess the confidence of each tracker's output and (b) a correlation scheme that evaluates the consistency of the trackers' output. The confidence and consistency metrics extracted are used to produce weights for the fusion of the available trackers.

1 Introduction

People tracking automation is becoming a cornerstone in security and surveillance applications, due to the enormous cost of human superintendence. As surveillance systems grow in scale, heterogeneity and capabilities, there is an increasingly critical need to provide an automated surveillance solution, able to perform under different conditions. Nevertheless, the existing tracking methods are still facing a number of challenges deriving from the diversity of the content that they have to process. The human mind can overcome these challenges imposed in multi-target people tracking by efficiently combining its inherent pattern recognition capabilities with semantically rich information gathered from the scene and the accumulated experience. In recent literature, multi-target tracking methodologies propose the combination of multiple types of trackers in a single system, in order to achieve robust tracking results. Given that each method has its own strengths and weaknesses, and operates optimally under different conditions, it is possible to create a general tracker with strong overall performance by combining different types of trackers in a *Late* or *Decision Level* fusion, where different cues are evaluated separately and the obtained decisions are fused.

Different late fusion methods have been presented, where multiple trackers are fused to produce more robust tracking results. In [1], [2], a multi-target, tracking-by-detection framework is proposed, which employs an hierarchy of trackers to select the most efficient tracking strategy. In [3], the authors combine a pedestrian detector and person specific classifiers in a particle filtering framework. In [4], authors describe a tracking-by-detection framework for multi-target tracking, including an occlusion reasoning stage. Other late fusion techniques include [5], where a number of weak labelers-trackers are fused through a majority voting scheme. Furthermore, in [6] the trackers are treated as black boxes. Their output consistency and correlation are calculated and then, they are fused using Gaussian Mixture Models. Late fusion schemes exploiting context awareness have also been proposed [7]. In [8], multiple characteristics of the scene are modelled to assist a tracker that segments the target in multiple blocks and follows a multi-level tracking scheme. However, motion based block tracking is illumination intolerant, which can prove problematic for scenes with rapid illumination change. In [9], an adaptive tracking algorithm is presented, combining probabilistic and deterministic trackers with a confidence estimation stage and pedestrians interactions to calculate reliable trajectories. However, its performance relies on a number of heuristics. Authors in [10] propose an off-line approach, where an incremental learning method of a non-linear motion map to produce more robust motion affinities between trajectories. In [11], an off-line method is described, where the social grouping behaviour of moving targets is modelled to assist the trajectory linking problem. The results are promising, however, the method seems likely to fail under heavy occlusions, where very close targets may merge.

Building on this approach, we present a novel multi-target tracking framework exploiting context information to enhance performance. A set of decision rules is created that implicitly integrates context-aware reasoning capabilities in multi-target scenarios. Rules take under account context information regarding the observed scene, status information for the targets being tracked, occlusion information in multi-target scenarios. The exploitation of context information and the enhanced occlusion reasoning method leads to a dynamically tuned multi-target solution for online tracking in surveillance videos.

The rest of the paper is organized as follows: after presenting the target profiling in Section 2, the baseline tracking scheme is provided in Section 3 and the multi-tracker fusion scheme is presented in 4. Experimental results are presented in Section 5 and finally, conclusions are drawn in Section 6.

2 Target Profiling

The Target Profiling subsystem is responsible to build an up-to-date profile for each target. The profile consists of appearance information, occlusion status and position.

2.1 Signature Extraction

The detector response is often imperfectly localized, including background segments. In order to produce a robust signature for each target, an automated segmentation process is applied to isolate the target, as described in [12].

Pedestrians often feature a bimodal color distribution in their appearance, deriving from the upper and lower clothing parts (i.e. blouse and trousers). A body division is applied to separate the templates into two parts, namely upper (I^u) and lower (I^l). It is proposed that the calculation of the division axis Y is performed by maximizing the Bhattacharyya distance d_{Bh} between the upper and lower HSV histograms $h(I)$, producing an accurate representation of the target.

$$argmax_y(d_{Bh}(h(I^u(y)), h(I^l(y)))) \quad (1)$$

The signature of each template consists of a combination of multiple color descriptors. HSV histograms and affine covariant regions (MSCR) are utilized .

2.2 Occlusion Status

We argue that a different tracking strategy should be followed for each one of the related targets according to its occlusion status. For that purpose, the notion *occlusion state* is introduced. Each target is classified to either independent or occlusion-related, based on the intersection of the bounding boxes (*bb*s). The intersection Q between the boxes bb_l and bb_m of people l, m, respectively, is defined as:

$$Q(bb_l, bb_m) = \frac{2 \cdot (bb_l \cap bb_m)}{bb_l \cup bb_m} \quad (2)$$

In order to have an occlusion and taking into account that *bb*s include background segments, an intersection of 30% of the total *bb* area has been heuristically defined as the threshold for occlusion, in all experiments.

An occluded object, depending on its relative position with other targets, can be an occluder, hiding other objects, or hidden by an occluder. In an occlusion scenario targets are labeled as occluders or occluded based on their similarity to their signature. The target with the minimum distance from its signature is categorized as the occluder. The rest are labelled as occluded. The occlusion state of each target is updated for every frame.

Occlusion-related targets are classified either as occluders or occluded. If a target's state is *occluder*, the tracking strategy remains the same. On the other hand, when the target's state is *occluded*, there is limited availability of appearance information. Thus, the classifier and the motion prediction module, during this time, halt their model updating, until the target becomes fully visible again.

2.2.1 Structure Modelling

The structure of the scene can be captured based on the statistical analysis of the motion detected in a training set. The pedestrian trajectories are extracted, filtered and, subsequently, used as input to automatically identify and label regions in the scene with certain characteristics. The output is a set of different masks, each mapping a different aspect of the scene structure, namely motion activity and entry/exit regions.

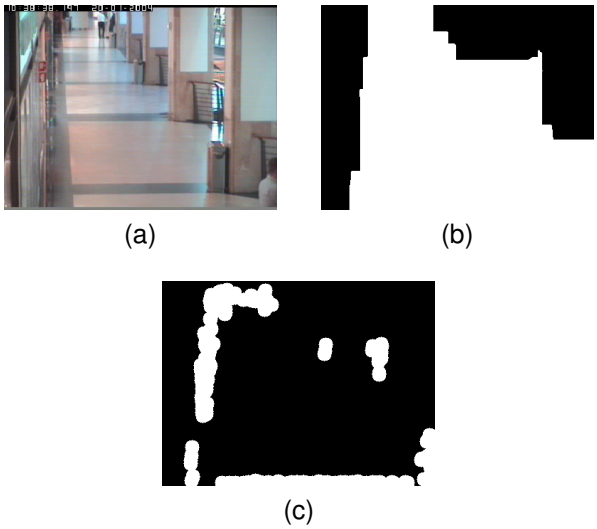

Figure 1. Examples of scene modeling masks from CAVIAR [13]. (a) Original frame, (b) Active areas mask, (c) Entry/Exit areas mask

The motion activity in each region of the scene is captured in the form of a binary mask, depicting areas with significant activity. To construct it, the coordinates of all the trajectories from the training set are collected and mapped on the scene. Morphological filtering is then applied resulting in a binary mask, where the $true$ area represents the active areas of the scene. The produced mask is denoted as *active area map* and an example is depicted in Fig. 1b.

Entry/exit zones are mapped using the starting and ending points of the accumulated trajectories gathered, excluding those identified in the first and last frames of the sequence because people may falsely appear to enter in the middle of the scene introducing noise. The qualified points are clustered into disjoint sets using the euclidean distance. Each identified cluster marks an entry/exit zone. Altogether, they form a convex hull. Assuming that a person cannot enter or exit from the inner area of the hull, clusters in the convex hull are excluded, while all remaining clusters form the *entry/exit zone map*. The final map is depicted in Fig. 1c as a binary mask.

3 Target Tracking

In this section, a methodology to fuse the tracking responses from multiple trackers through a dynamic context-aware process to improve tracking efficiency is presented.

3.1 Target Initialization/Termination

Due to the detector's uncertainty, initialization/termination of the targets when they enter/leave the scene, can be challenging. An accumulative voting system is proposed to robustly identify the entrance/exit of people in the scene, overcoming the detector's sparsity and false positive detections. When a new target is detected, a hypothesis is formulated. If enough consecutive detections are accumulated, the hypothesis is accepted and it is assigned a unique identification number (ID). The initialization threshold th_{init} is defined based on the appearance frequency of the detector's false positive detections (FPs). A threshold is defined, by requiring at least 99% of the FPs to be eliminated. This requirement removes possible outliers that will affect the threshold definition. To terminate tracking, when the object exits the scene, a similar voting scheme is employed, counting frames with no matching response. When a threshold th_{term} is reached, the target is terminated.

The thresholds calculated using the statistical processing of the detector responses are generally valid for the scene. However, the tracker's performance can be improved by modifying them in a region level according to its motion activity and whether it consists entry/exit zones. In regions with no motion activity, the appearance of a new target is rare and, therefore, the initialization threshold is increased to avoid false detections. The termination threshold in the same area is decreased to assist the quicker deletion of false target detections or valid targets that have drifted away. On the other hand, in entry/exit zones the thresholds for initialization and termination are both lowered as the possibility of a change in target population number is very high. A statistical analysis similar to the one described above was performed for each of the respective areas.

Moreover, a validation process for each newly identified target is proposed to avoid the re-initialization of targets whose tracking has been lost due to extensive occlusion events. The signature extracted from each initialized target is compared with the signatures of the existing targets to identify possible matches. In order to minimize the number of possible candidates, the pool of possible matches is restricted to targets in the vicinity that are labeled as occluded. If the newly initialized target matches a prior target it regains its former ID. Otherwise, the target is considered new. This module is activated in the proposed framework on-demand, when a new object appears, to minimize the computational burden.

3.2 Trackers Pool

The proposed fusion framework is tracker-agnostic. In order to test it, a pool of state-of-the-art trackers has been assembled, containing a detection-based tracker, a target-specific classification-based tracker and an appearance-independent motion predictor.

Tracking by Detection: Object detection can facilitate a basic multi-target tracker with robust performance in simple cases. Current detectors combine speed, performance and invariance to considerable changes in lightning and scale. By applying the detector on every frame, we construct a sequence of detection response maps depicting possible object localizations. A detector can localize many instances of the class in a frame, allowing multi-target tracking. The tracking of the detected objects is accomplished by linking their responses along consecutive frames, in an association process commonly known as the assignment problem.

A detector requires target visibility and heavily relies on the training of the model itself. The detection response can be either sparse or include a lot of false positives. A low detection rate might lead to undesirable side effects such as tracking inaccuracies or an untimely tracking termination. Moreover, when two targets collide (occlusion), the detector merges the targets in one response, limiting the accuracy of this approach in scenarios with occlusions.

Tracking by Classification: A tracker needs to know a priori each target's appearance independently to resolve association ambiguities in multi-target and occlusion scenarios. A target-specific classifier can localize a specific object in subsequent frames, given an initial template. The model is updated with a predefined learning rate λ that limits the contribution of new templates to the classification model. The optimal learning rate value depends on the content and the characteristics of the video. A classifier is initialized for each one of the targets to model their appearance in the scene.

While online adaptation of the classification model is essential to track non-rigid and variable targets (e.g. a person), it introduces a gradual failure of the template a.k.a. drifting. Due to imperfect target localization and scale changes, background features are increasingly incorporated in the template, leading to a gradual template failure and, finally, the loss of the target.

Tracking by Motion prediction: A motion prediction approach that is appearance-independent is proposed in this work, based on [14]. The proposed approach exploits the observed motion behavior of targets on the scene, and builds on the local motion models acquired through motion modeling to create an online motion prediction module.

The first step in motion prediction is to create the motion model for the examined scene, based on prior motion patterns. The accumulated prior trajectories that are used as training material are divided into smaller tracklets with a fixed length $N_{tracklet}$. The informative tracklets are filtered, producing a large set that summarize the motion patterns observed in the scene. The dominant tracklets in each neighbourhood represent local motion models. Gaussian Process (GP) regression is used in order to model the dominant motion patterns.

The local motion models identified are exploited to create an online motion prediction module. Given a person in the scene, a tracklet containing the $N_{tracklet}$ prior locations of the target is fed to the Motion Prediction module. This tracklet is assigned to a grid point of the scene, based on its localization. The motion models that correspond to this grid point are employed to estimate the next position of the target.

4 Tracker Fusion

In this section, a context-aware fusion of independent trackers is introduced for multi-target tracking in surveillance video sequences. A late fusion scheme is employed, fused with context information to streamline performance. In order to fuse different trackers efficiently, a confidence metric of their response is required. However, trackers do not produce any metric or they produce metrics which are not comparable. Therefore, a tracker-agnostic framework to assess their performance is presented.

4.1 Tracker Confidence

In order to facilitate the evaluation of the tracker confidence, information cues that measure the robustness of the tracker are proposed, capturing the consistency of the tracking responses, in terms of speed and appearance.

Speed: In a small time window, it can be assumed that objects are moving in a stable speed (inertia). Therefore, a sudden change could be an indication of a tracking failure. Speed similarity is defined as:

$$Sim_{sp} = \frac{min(\bar{V_N}, V_{T_k})}{max(\bar{V_N}, V_{T_k})} \tag{3}$$

where $\bar{V_N}$ is the mean speed in the last N frames and V_{T_k} is the speed in the examined frame.

Appearance: In the same time window, appearance information is also likely to be preserved, fully or partly. Therefore, a sudden change in the appearance of a moving target could also be an indication of a tracking failure. In order to test the tracking consistency, the candidate response is compared against prior observations, using appearance signatures to extract an appearance similarity score Sim_{app}.

We assume that a candidate response with consistent speed and appearance has a significant probability to be correct. Deviation in either one could be an indication of a tracking failure. Therefore, the overall confidence S_i of tracker i, normalized to $[0, 1]$, is defined as:

$$S_i = \frac{Sim_{sp} + Sim_{app}}{2} \tag{4}$$

4.2 Tracker Accordance

In a frame, every individual tracker provides the system with a different tracking response. Besides the performance of each tracker, the pair-wise proximity of the respective responses is also an indication of accurate tracking. Response similarity is calculated taking under account: distance, overlap and resemblance of the responses, as presented below.

Distance: Two trackers do not produce exactly the same responses for the same target. Nonetheless, the centers of the responses should be close, to increase their credibility. The similarity Sim_{dist} of two responses with bounding boxes bb_i and bb_j, in terms of their distance, is defined as the euclidean distance of their centers.

Overlap: Besides the distance between two responses, the size and shape of their bounding boxes contain essential information. Therefore, the overlap of two responses is also employed as a similarity measure Sim_{ovrl}.

Appearance: Appearance similarity among tracker responses can be utilised also to assess the pair-wise tracker accordance. The appearance similarity (Sim_{app}) among the responses is employed to evaluate the appearance variation among them.

We assume that pairs of tracking responses with high similarities in terms of distance, overlap and resemblance have a significant probability to be accurate. Therefore, the overall Accordance metric $A_{i,j}$ between trackers i, j is defined as:

$$A_{i,j} = \frac{Sim_{dist} + Sim_{ovrl} + Sim_{app}}{3} \tag{5}$$

where Sim_{dist}, Sim_{ovrl}, Sim_{app} are the distance, overlap and appearance cues, respectively.

4.3 Dynamic Fusion

A fusion scheme that adjusts to the confidence and accordance of the trackers' responses is proposed here. Intuitively, candidate responses with high confidence, whose responses are confirmed by competing trackers, are more likely to represent the true position of the target and therefore, their contribution leads to more accurate tracking results. It is important, thus, to employ a fusion scheme which will "reward" the reliable responses, while it will "punish" the weak ones.

Towards this end a pair-wise combinatorial fusion is employed. Trackers are dynamically combined in pairs to produce intermediate results using the tracker's confidence. The intermediate results are then fused using the tracker accordance metric. As a result, responses that share common characteristics are boosted during fusion. The final response can be calculated as:

$$R = \sum \frac{A_{i,j}}{A_{sum}} \cdot (\frac{S_i}{S_i + S_j} \cdot R_i + \frac{S_j}{S_i + S_j} \cdot R_j) \tag{6}$$

where R_k is the response of tracker T_k.

5 Experimental Results

The proposed approach is applied for evaluation purposes on two different publicly available datasets: CAVIAR and PETS 2009, which have been widely used as tracking evaluation datasets in the literature.

Two sets of metrics are used to cover all aspects of tracking evaluation. The first set is the CLEAR MOT metrics, which focuses on the frame to frame performance of the tracker. The second set of metrics used is the trajectory quality metrics, which focus on performance in the trajectory level. The trajectories are classified to mostly tracked (MT), partially tracked (PT), and mostly lost (ML), which are trajectories successfully tracked for more than 80%, more than 20%, and less than 20%, respectively. Moreover, fragmentations (FM) and identity switches (IDS) are computed.

5.1 Evaluation of fusion contribution

In this section, the contribution of the trackers and the dynamic fusion in the proposed system is evaluated and analyzed. We use the CAVIAR dataset for the evaluation due to its more complicated scenarios and its total frame length, and the results are given in table 1. First, the results for each primary tracker, the naive fusion methodology and the proposed one are presented.

Tracker	MT	PT	ML	FM	IDS
Detector	83.33%	13.89%	2.78%	49	31
Classifier	43.06%	43.75%	13.2%	83	32
Predictor	6.25%	59.03%	34.7%	60	7
Naive Fusion	89.58%	9.722%	0.694%	40	25
Proposed	91.70%	7.60%	0.70%	29	13

Table 1. Evaluation of module contribution on CAVIAR dataset

By observing the results, it is obvious that the detector performs better than the other two. It is also clear, that the classifier and the predictor cannot stand as individual trackers, something that was expected due to their nature. A naive fusion of all trackers, where their mean value is used already improves MT trajectories and decreases FMs and $IDSs$.

Albeit the increase in the MT by the naive fusion, still the number of fragmentations and ID switches is high. These errors are caused mostly during occlusions, where the detector and the classifier cannot cope with the appearance changes. Our context aware novel dynamic fusion scheme aims to boost strong tracking responses with common characteristics and reduce the effect of weaker tracking responses to the final result. The results affirm the tracking improvement, since the MT has been further increased, and at the same time, the number of FM and IDS has been significantly decreased.

5.2 Comparison with SoA

In this section, a comparison of the proposed framework against other state-of-the-art online multi-target trackers is presented. The three datasets were employed and the evaluation was performed using the parameters summarized in Table 2. In order to render our results comparable with the related literature, the detection responses presented in [15] are used. The proposed framework is compared against online trackers presented in Zhang et.al. [16] and Duan et al. [8] and offline trackers presented in Kuo et al. [15], Yang et al. [17], and Nevatia et al. [10].

Parameter	CAVIAR	PETS09
$frame$	384x288	768x576
fps	25	7
th_{init}	7	3
th_{term}	11(5)	8(3)
$learn_{rate}$	0.15	0.35

Table 2. Parameter set for the evaluated datasets.

Tracker	MT	PT	ML	FM	IDS
Kuo et al.[15]	84.6%	14.7%	0.7%	18	11
Nevatia et al.[10]	89.1%	10.2%	0.7%	11	5
Duan et al.[8]	89.7%	7.4%	2.9%	29	15
Proposed	91.7%	7.6%	0.7%	29	13

Table 3. CAVIAR statistical results using the trajectory quality metrics

Tracker	MT	PT	ML	FM	IDS
CVPR10 [15]	82.6%	17.4%	0.0%	21	15
CVPR11 [17]	78.9%	21.1%	0.0%	23	1
CVPR12 [10]	89.5%	10.5%	0.0%	9	0
AVSS12 [16]	78.9%	15.8%	5.3%	15	5
Proposed	94.7%	5.3%	0.0%	34	7

Table 4. PETS 2009 S2L1 (view 1) sequence statistical results using the trajectory quality metrics

6 Conclusions

In this work, an online multi-target multi-tracker fusion framework is presented, infused with context information to create an on-line tracker. The contribution of the proposed work lies in the exploitation of accumulated context information to assist the tracking procedure. Semantic information regarding entry/exit zones, activity areas, and target modeling are employed to assist tracking by guiding the association process and enabling adaptive, region-level parameter setting. Furthermore, the inter-object relations are investigated and they are exploited to resolve occlusion related issues. The responses of the tracker pool are fused using a dynamic fusion strategy, based on a number of visual and spatio-temporal coherence rules.

The experimental evaluation on three different datasets shows that the proposed framework provides very promising results. Nevertheless, motion prediction of the targets is not always accurate and deviations from the real trajectory are observed. Predicting the human motion is not always possible but it could be improved using a continuously updated motion model that will feature ever more trajectory examples. Another aspect of the architecture employed is the selection of the trackers. It is important to opt for trackers that cover different aspects of the tracking procedure, in order to complement each other and balance their weaknesses. Thus, the final system will be more robust and will efficiently adapt to different tracking scenarios.

Acknowledgements

This work was supported by the European Community's funded project LASIE (www.lasie-project.eu) under Grant Agreement no. 607480.

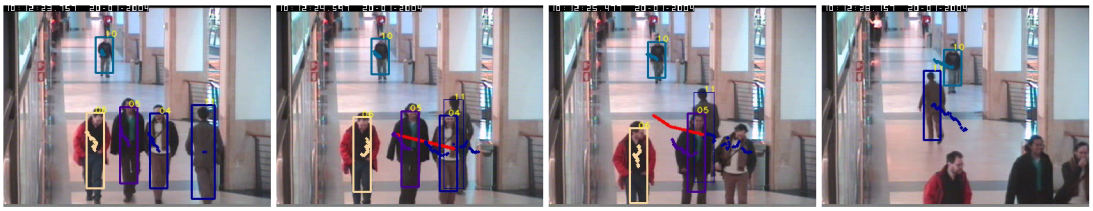

Figure 2. Results for the CAVIAR dataset.

Figure 3. Results for the PETS 2009 dataset.

References

[1] J. Zhang, L. L. Presti, and S. Sclaroff, "Online multi-person tracking by tracker hierarchy," in *Ninth International Conference on Advanced Video and Signal-Based Surveillance*, AVSS '12, pp. 379–385, IEEE, 2012.

[2] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler, "Re-identification of pedestrians in crowds using dynamic time warping," in *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pp. 423–432, Springer, 2012.

[3] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 9, pp. 1820–1833, 2011.

[4] N. McLaughlin, J. M. del Rincon, and P. Miller, "Online multiperson tracking with occlusion reasoning and unsupervised track motion model," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 37–42, 2013.

[5] B. Zhong, H. Yao, S. Chen, R. Ji, T.-J. Chin, and H. Wang, "Visual tracking via weakly supervised learning from multiple imperfect oracles," *Pattern Recognition*, vol. 47, no. 3, pp. 1395 – 1410, 2014. Handwriting Recognition and other {PR} Applications.

[6] Y. Gao, R. Ji, L. Zhang, and A. Hauptmann, "Symbiotic tracker ensemble toward a unified tracking framework," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 24, pp. 1122–1131, July 2014.

[7] L. Snidaro, J. Garca, and J. Llinas, "Context-based information fusion: A survey and discussion," *Information Fusion*, vol. 25, no. 0, pp. 16 – 31, 2015.

[8] G. Duan, H. Ai, J. Xing, S. Cao, and S. Lao, "Scene aware detection and block assignment tracking in crowded scenes," *Image Vision Comput.*, vol. 30, pp. 292–305, May 2012.

[9] A. Bera, D. Manocha, A. Lake, and N. Galoppo, "Adapt: Real-time adaptive pedestrian tracking for crowded scenes," in *Robotics and Automation (ICRA), International Conference on*, pp. 1801–1808, IEEE, 2014.

[10] R. Nevatia and Y. Bo, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *Conference on Computer Vision and Pattern Recognition*, IEEE, 2012.

[11] Z. Qin and C. R. Shelton, "Improving multi-target tracking via social grouping," in *CVPR*, pp. 1972–1978, 2012.

[12] V. Lovatsis, A. Dimou, and P. Daras, "Introducing context awareness in multi-target tracking using re-identification methodologies," in *ICDP*, 2013.

[13] "Caviar dataset (2nd set, corridor view)," 2003.

[14] V. Lasdas, R. Timofte, and L. Van Gool, "Non-parametric motion-priors for flow understanding," in *Workshop on the Applications of Computer Vision*, IEEE, 2012.

[15] C.-H. Kuo, C. Huang, and R. Nevatia, "Multi-target tracking by online learned discriminative appearance models," in *Computer Vision and Pattern Recognition, IEEE Conference on*, pp. 685–692, IEEE, 2010.

[16] J. Zhang, L. L. Presti, and S. Sclaroff, "Online multi-person tracking by tracker hierarchy," *Advanced Video and Signal Based Surveillance*, vol. 0, pp. 379–385, 2012.

[17] B. Yang, C. Huang, and R. Nevatia, "Learning affinities and dependencies for multi-target tracking using a crf model," in *Computer Vision and Pattern Recognition, IEEE Conference on*, pp. 1233–1240, IEEE, 2011.