# Introducing Context Awareness in Multi-Target Tracking using Re-Identification Methodologies

**V. Lovatsis, A. Dimou and P. Daras**

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

**Keywords:** surveillance, tracking, re-identification, occlusion reasoning.

## Abstract

In this paper, re-identification techniques are exploited to add context awareness to a multi-target tracker and enhance its tracking performance, in an online manner. To achieve that, targets are labeled as independent, occluders or occluded ones, based on the completeness of their appearance information. For each category, a different tracking strategy is employed to achieve the optimal results. In cases of tracking failure, an online automated re-identification technique is proposed, to alleviate multiple identity assignments to the same target. Experimental evaluation conducted on the CAVIAR and PETS 2009 datasets shows that the proposed mechanism enhances tracking performance compared to a baseline tracker and achieves competitive performance with state of the art methods.

## 1 Introduction

Surveillance automation becomes increasingly important as new cameras are installed daily in public or private areas. Automation aims at real-time processing of aggregated footage, reducing human effort and interaction. Online tracking of multiple objects in semi-crowded environments is a very active research area in computer vision [1, 2, 3]. A tracker has to overcome challenges like appearance variations [4], intra-class discrimination [5], scene occlusions [6] and combinations of the above. Tracking can also be expanded into a network of cameras where targets are associated along different cameras [7].

Detection-based trackers have emerged as a popular choice for tracking due to their improved performance and accuracy [8]. They are using appearance models to find the new position of an object by re-detecting it in every frame. The tracker's observation model is based on trained detectors capable of localizing an object class (pedestrians, cars, etc.) in many scales and viewing angles [8, 9]. The association of the detection responses across frames is based on spatiotemporal constraints. To further improve discriminability between targets of the same class during inter-object occlusions, the use of adaptive, target-specific classifiers, assigned on every object, has been widely proposed. The appearance models in these methods aim at making the targets of the same class distinguishable [10]. Recently, a new trend has emerged where the track by detection-based tracker is supported with target-specific classifiers. The combined responses of the two trackers are producing a more robust tracking either for single-target [11], or multi-target scenarios [12].

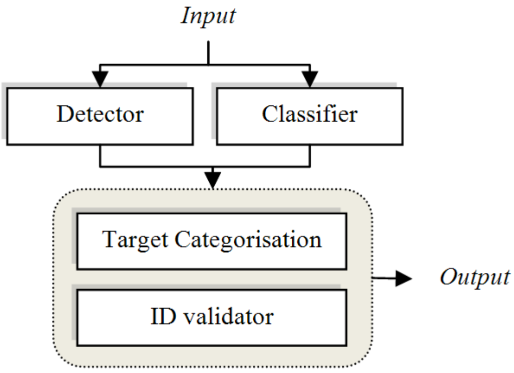


Figure 1. Proposed mechanism. Target categorization sets the fusion rules of the observation models while ID validation checks the originality of newly found objects.

Multi-target trackers can be categorized to online and offline based on the association methodology followed and the use of post-processing techniques. Online methods consider only past and current frames and take decisions in a causal way [1]. Offline methods get information from future frames as well, and they use post-process data, such as energy minimization schemes [13], to link fragmented trajectories.

Person re-identification (Re-ID) is being studied intensively but mainly for inter-camera target associations [14, 15, 16, 17]. The problem is approached by appearance modeling, where the need for robust descriptors becomes a priority [15]. There are few works that combine Re-ID with intra-camera tracking. In [18], a matching method is used in combination with a motion-based tracker. The work in [19] associates tracklets in a sliding window, where linking is achieved without intermediate knowledge between matching frames or any form of data pre-processing, while [20] associates trajectories in an empirical way. In most of the related work Re-ID is employed in post-processing schemes to associate tracklet, while in the presented work it is used for context awareness in an online fashion.

In this paper, the main contribution is the introduction of a context-aware target-labeling procedure, using Re-ID techniques to enable dynamic tuning of the tracking parameters. Each target is labeled based on the completeness of its appearance information as independent, occluder or occluded. A baseline tracker, constructed combining a detector and a classifier, is enhanced with the proposed context-aware labeling. For each category, it employs different fusion weights for the detector and the classifier responses. Moreover, Re-ID is used to prevent targets from acquiring multiple identities. Every time that a new target is introduced close to an occlusion area, it would normally acquire a new identity number (ID). An automated Re-ID based technique is proposed to validate newly found targets and link them with their previous IDs when needed. The proposed approach runs in a causal way and association decisions are taken online upon each frame.

The rest of the paper is organized as follows: the baseline tracker is described in Section 2, while intra-camera Re-ID is analysed in Section 3. The proposed method is described in Section 4. Experimental results on public benchmark datasets are evaluated in Section 5, followed by the conclusions in Section 6.

## 2 Baseline Tracker

A baseline tracker has been implemented as a tracking reference to validate the advantages of the proposed attachable module. Initially, an object detector is applied on each frame to construct a response map of object localizations. Association of the objects across frames is, then, facilitated by an online learning-based classifier. The classification unit is adapted through time on each target appearance. Association is achieved based on the assumption that objects cannot move drastically between consequent frames. In our experiments, the baseline tracker used the detection responses reported in [11] for the pedestrian detection and the Compressive Tracker [21] as the target-specific classifier.

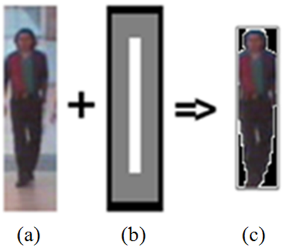


Figure 2. The segmentation of a cropped frame (a) is guided by a trimap (b), color-mapped with foreground pixels in white, probably foreground pixels in gray and background pixels in black. As a result, target (c) is represented only by relevant pixels.

Detection responses are depicted as rectangular regions, where each region is described by a bounding box. For each successive frame, the correspondence process attempts to associate each region with one of the existing tracks based on spatiotemporal constraints [5]. Single-object tracking is feasible with a detection-based association. However, in multi-target tracking scenarios, association becomes challenging, because a single response can belong to many objects during an occlusion, causing ambiguities. Therefore, a tracker must exploit each target appearance individually to increase its intra-class discrimination capability by using a classifier on each target. Moreover, detection responses can be sparse resulting in tracking errors.

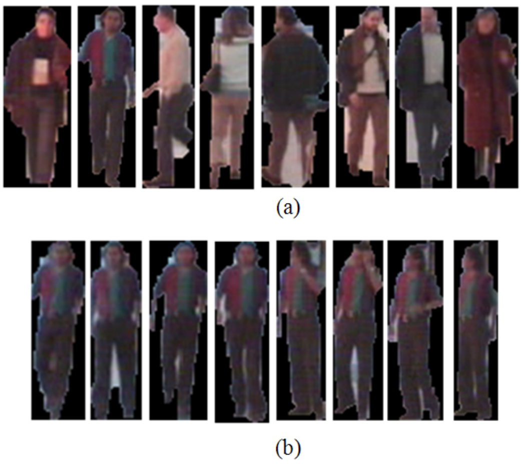


Figure 3. (a) Unsupervised segmentation shots extracted from one frame. (b) Target shots are accumulated every Nth frame into the gallery to construct a multi-shot representation.

The tracker is initialized by each new entry found by the detector. Using two sets of samples, characterizing the inner box area as positive and the outer space as negative, Haar-like features are extracted and used to train a naive Bayes classifier [21]. During tracking, the model is being updated according to a predefined learning rate so as to optimize tracklets' construction. During classification, the search window is set to cover the area around the target. Target-specific adaptation minimizes object's centre variation, guides tracker when detector response is sparse and resolves short-term occlusions successfully. Update is suspended when objects are participating in an occlusion. However, classifier's performance is based on its limited search window size which is typically appropriate for localizing an object on successive frames. The weakness of classifier's search method is revealed when objects are either occluded for a long period of time or the template undergoes large appearance differences , resulting in the template update problem (a.k.a. drifting) [22].

Detector's and classifier's predictions, $DP_k^i$ and $CP_k^i$ respectively, can compensate each other's errors by using a weighting scheme. Detector's sparsity and its inability to handle short-term occlusions and the classifier's drifting problem can be overcome by fusing both predictions into a final one $FP_k^i$ using a weight function:

$$FP_k^i = \boldsymbol{w} \cdot \begin{bmatrix} DP_k^i \\ CP_k^i \end{bmatrix} \quad (1)$$

where $\boldsymbol{w} = [w_d \;\; w_c]$ is the weights vector. As a result, the detector acts as the unbiased observation model while the classifier refines results in an adaptive way. The default fusion weights of the detector and the classifier predictions are equiponderant. In section 4.1, a mechanism is proposed to configure the weights according to the scene context.

## 3 Intra-camera Re-ID

In the relevant literature, Re-ID methods are considered independent from tracking and are tested on datasets with already cropped targets accompanied with Ground Truth annotation for segmenting foreground pixels from the background [16]. In an autonomous tracking framework, though, a preprocess unit is required to construct the image gallery. Also, Re-ID matching relies on target's appearance, thus object segmentation is necessary to construct representative descriptors. However, most trackers provide detection windows around the objects, where the localization may not be centered in the box or the box may not have the correct size.

There is a plethora of methods for segmenting a target inside a window [23, 24, 25]. There have been impressive developments in techniques of semi-automatic segmentation, where user interactions refine the results [26]. We propose to fully automate segmentation based on the assumption that, despite the inaccurate localization of the detector, the center of tracker's detection window usually contains the most relevant information. Therefore, segmentation is biased to accept pixels in the center of the box and reject the ones in the boundaries.

The pre-process unit is fed with cropped boxes of independent objects in parallel with tracking. Segmentation is guided by an initial trimap $T = T_F, T_{PF}, T_B$, tagged with areas for foreground $T_F$, probably foreground $T_{PF}$ and background pixels $T_B$ (Figure 2). For efficiency, the number of shots per representation is limited. To counterbalance appearance changes through time, shots are aggregated in each Nth frame. Consequently, every tracklet can be represented by a volume of cropped and segmented images, i.e. a multi-shot representation. The representations are accumulated to construct the gallery. Results of this process can be seen in Figure 3.

Re-ID approaches consider subjects as sets of local and global features extracted from a set of images [15]. Asymmetry driven body division is applied to separate the box into body parts by maximizing the difference between upper and lower HSV histograms of the human body under the assumption that pedestrians have a bimodal chromatic distribution (i.e. blouse and trousers). Along with the histograms, the Maximally Stable Color Regions [27], that encode the texture information, are accumulated into the representation. Derived either from a single frame or from a sequence of frames, a single or a multiple signature is generated. Matching of the signatures produces ranked results where the first ranking position indicates the best pair for linkage. Matching is based on appearance similarity and is expressed by the distance $d_{ReID}$ between the target $A$ and the matching candidate $B$:

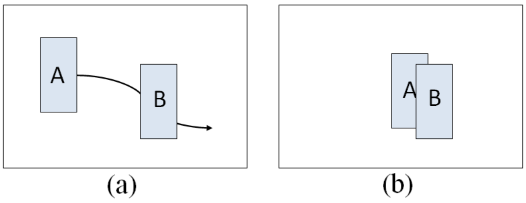
$$d_{ReID}(A, B) = d_{HSV}(A, B) + d_{MSCR}(A, B) \qquad (2)$$

| Labels/Weights | $w_d$ | $w_c$ |
|---|---|---|
| Independent | 0.5 | 0.5 |
| Occluder | 0.2 | 0.8 |
| Occluded | 0.8 | 0.2 |

Table 1. Fusion weights of trackers based on object state.

where $d_{HSV}$ is the Bhattacharyya distance between the target's respective parts (upper and lower) and $d_{MSCR}$ is the MSCR distance [27]. The aforementioned distance regards only the case where two single-shot representations are compared. Our technique falls within the case of Single-shot vs Multi-shot signature matching [14]. To compare a single-shot representation versus a multi-shot one, the mean distance is used, derived from each comparison between the single-shot representation with each shot from the multi-shot one.

For the Re-ID module, the methodology presented in [14] was ported in C++ to be integrated to our framework. The texture descriptor in the original version was excluded due to its disproportional computational burden. For independent targets, a tracklet of 3 frameshots per second was produced, regardless the video frame rate, and a total of 5 frameshots were accumulated into each multi-shot representation based on the observation reported in [16]. For object segmentation on the frameshots, we used [24] with the user-defined mask accepting central points as foreground elements.



(a) (b)

Figure 4. In (a), independent objects $A$ and $B$ are about to meet. In (b), occlusion has begun with $A$ passing behind $B$.

## 4 Occlusion Handling

A tracked object can be either fully independent from other objects, or an occluder hiding other objects or hidden by an occluder. In the proposed framework, it is argued that a different tracking strategy should be followed in each case. Independent objects are easy to track since there is no noticeable detection sparsity. Problems occur when objects get cluttered with each other. Inter-object occlusions are detected when their bounding boxes overlap (Figure 4).

### 4.1 Target Categorization

In order to categorize the targets, the completeness of their appearance is evaluated. During an occlusion, the content of an occluded bounding box might miss valuable appearance information and only the most-front object has full visual completeness. Based on the inherent capability of the re-identification technique to order objects based on their appearance similarity, targets are labeled as occluders or occluded.

Assume $K$ objects participating in an occlusion at frame $n$. To categorize the bounding boxes during frame $n$, each box is compared against its multi-shot representation, extracted as described in Section 3. Each comparison, produces a similarity score $d_{ReID}$ between the box's single-shot at frame $n$ and the multi-shot representation since the object was independent before frame $n-1$. A total of $K$ comparisons are produced and the target with the minimum distance is categorized as the occluder. The rest $K-1$ boxes are labelled as occluded.

At this point, the baseline tracker is aware of the target categorization and can now use different strategies for each target group. For occluders, fusion favors the classifier, while for hidden targets, fusion favors the detector since the template might not be visible. The different weights for each target category can be seen in Table 1.

### 4.2 ID Validation

A common tracking error occurs when an object gains multiple ID numbers. There are cases where long-term occluded targets are lost and may re-appear outside of the search window of the classifier resulting into the initialization of a new target. Their original bounding boxes cannot be terminated due to the persistency of occluded bounding boxes. The tracker terminates only independent tracklets in the absence of detection responses. To prevent ID switching errors and the drifting of orphan occluded boxes, the Re-ID mechanism is triggered to check the originality of newly found targets near occlusions.

Any attempt for a new entry $(FP^x)$ by the tracker, around an occluded area $S$, must be compared against all candidates classified as occluded. The minimum of the comparison distances $(d_{ReID}^{min})$ indicates the best pair for association. Due to the fact that results are ranked, a global threshold $th_{ReID}$ is used to ensure that the minimum distance is within acceptable limits. If $d_{ReID}^{min}(FP^x, \{FP^i \epsilon S\}) < th_{ReID}$ , the target regains its ID number. Otherwise, a new target is initialized.

In literature, Re-ID methods use distance metrics for pairwise scoring leading to efficiency and simplicity. Nevertheless, no absolute confidence measure exists due to the absence of machine learning algorithms during the feature extraction stage and the fact that association is treated as a ranking problem. This leads to the inevitable use of thresholds, which are heuristically defined in order to maximize true matches [16]. During our experiments, a threshold of $th_{ReID} = 0.7$ is used in all cases. In order to define this threshold, a bimodal distribution of the correct and wrong matching distances was created. By fitting two Gaussian distributions on the distances data and identifying their intersection, the threshold is set, as in [14].

## 5 Experimental Results

The proposed approach is applied for evaluation purposes on two different datasets: CAVIAR[1] and PETS 2009[2], which have been widely used as tracking evaluation datasets in literature. The CAVIAR project dataset depicts the view across a hallway in a shopping center. The ground plane that stretches among the z-axis in combination with the low positioning of the camera results in sequences with many long-term occlusions. In total, it includes 26 video sequences, containing a varying number of individuals and groups. However, 20 of the 26 sequences were used as testing set as in [10]. The average length of the video sequences is 1500 frames. The resolution of the frames is 384 x 288 pixels and the frame rate of each sequence is at 25 frames per second (fps). Following the literature [3], the ground truth data of the CAVIAR dataset is filtered, removing objects that are too small or partially out of the scene.

| Tracker | Type | GT | MT | PT | ML |
|---|---|---|---|---|---|
| Li et al. [10] | Offline | 143 | 84.6% | 14.0% | 1.4% |
| Bak et al. [19] | S.window | 140 | 84.6% | 9.5% | 5.9% |
| Baseline tracker | Online | 138 | 85.5% | 12.3% | 2.2% |
| Proposed method | Online | 138 | 86.3% | 10.1% | 3.6% |

Table 2. Comparison of different tracking results on the CAVIAR Dataset.

| Tracker | Type | GT | MT | PT | ML |
|---|---|---|---|---|---|
| Zhang et al. [12] | Online | 19 | 78.9% | 15.8% | 5.3% |
| Badie et al. [20] | Offline | 12 | 50.0% | 33.3% | 16.0% |
| Baseline tracker | Online | 19 | 73.7% | 26.3% | 0% |
| Proposed method | Online | 19 | 78.9% | 21.1% | 0% |

Table 3. Comparison of different tracking results on the PETS 2009 S2L1 View 01 sequence.

From the PETS 2009 dataset, the sequence S2L1 is used for the experiments. It depicts a campus road where a sparse crowd is walking. Challenges for this sequence include the low sampling rate of the camera which produces fast moving objects and multiple occlusions. Moreover, a sign at the center of the frame constitutes a scene occluder which covers all people behind it and in some cases for long periods. The length of the video sequence is 795 frames and the resolution of the frames is 768 x 576. The frame rate of the sequence is at 7 frames per second.

For comparison with other state-of-art methods, we adopted commonly used metrics [28]. The metrics used are:

- GT: the number of ground truth trajectories.
- MT: the percentage of trajectories successfully tracked for more than 80% of their total length.
- PT: the percentage of trajectories that are tracked between 20% and 80% of their total length.
- ML: the percentage of trajectories that are tracked for less than 20% of their total length.

The higher value, is better for MT, while the lower value, is better for PT and ML.

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1

[2] http://www.cvg.rdg.ac.uk/PETS2009/

Tracking evaluation results are depicted in Tables 2 and 3. For the Caviar dataset, the comparison shows that the proposed framework achieves the most MT, while keeping the ML lower than [19]. Compared to the offline method [10], our system achieves the most MT, but more ML trajectories. Offline trackers consider all detection responses across frames given and can handle detection sparsity better than the online approaches that decide in a causal way. For the PETS 2009 dataset, the comparison shows that our system outperforms all approaches, achieving the same MT as [12], while ML is 0. These results show that the baseline tracker produces comparable results considering the state-of-the-art methods, while our attachable mechanism enhances and improves the overall performance of the proposed framework. Example sequences of output frames are depicted in Figures 5 and 6.

## 6 Conclusions

In this work the exploitation of Re-ID techniques is proposed to enhance a multi-target tracker, introducing dynamic parametrization of the tracker on a target-based level. The proposed methodology labels the targets according to their appearance completeness to individual, occluder or occluded, employing a different tracking strategy in each case. Moreover, the proposed framework is exploiting Re-ID to alleviate multiple ID assignment to the same target. It must be noted that all the above functionalities are available in a online fashion and they could be integrated to any multi-modal tracker. The results of the experimental evaluation show a significant improvement, compared to the baseline tracker and state-of-the-art tracking methods employing Re-ID techniques.

## Acknowledgements

This work was supported by the European Community's funded project ADVISE (www.advise-project.eu) under Grant Agreement no. 285024.

## References

[1] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *PAMI*, vol. 33, no. 9, pp. 1820–1833, 2011.

[2] W. Hu, X. Li, W. Luo, X. Zhang, S. Maybank, and Z. Zhang, "Single and multiple object tracking using log-euclidean riemannian subspace and block-division appearance model," 2012.

[3] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *ECCV 2008*, Springer.

[4] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *Signal Processing, IEEE Transactions on*, vol. 50, no. 2, pp. 174–188, 2002.

[5] A. Senior, A. Hampapur, Y. Tian, L. Brown, S. Pankanti, and R. Bolle, "Appearance models for occlusion handling," in *International Workshop on PETS*, 2001.

[6] B. Babenko, M. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *CVPR*, 2009.

[7] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *CVPR*, vol. 2, pp. 26–33, IEEE, 2005.

[8] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *PAMI*, vol. 34, no. 4, pp. 743–761, 2012.

[9] J. Wu, C. Geyer, and J. Rehe, "Real-time human detection using contour cues," in *Robotics and Automation (ICRA)*, pp. 860–867, IEEE, 2011.

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

[11] C. Kuo, C. Huang, and R. Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," in *CVPR*, IEEE, 2010.

[12] J. Zhang, L. Presti, and S. Sclaroff, "Online multi-person tracking by tracker hierarchy," *Advanced Video and Signal Based Surveillance, IEEE Conference on*, 2012.

[13] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *CVPR*, IEEE, 2008.

[14] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, IEEE, 2010.

[15] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, Springer, 2008.

[16] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person re-identification," *PAMI*, 2012.

[17] D. N. T. Cong, L. Khoudour, C. Achard, C. Meurie, and O. Lezoray, "People re-identification by spectral classification of silhouettes," *Signal Processing*, 2010.

[18] A. Romero, M. Gouiffes, and L. Lacassagne, "Covariance descriptor multiple object tracking and re-identification with colorspace evaluation," *ACCV*.

[19] S. Bak, D. Chau, J. Badie, E. Corvee, F. Brémond, and M. Thonnat, "Multi-target tracking by discriminative analysis on riemannian manifold," in *ICIP*, IEEE, 2012.

[20] J. Badie, S. Bak, S. Serban, and F. Bremond, "Recovering people tracking errors using enhanced covariance-based signatures," in *AVSS*, IEEE, 2012.

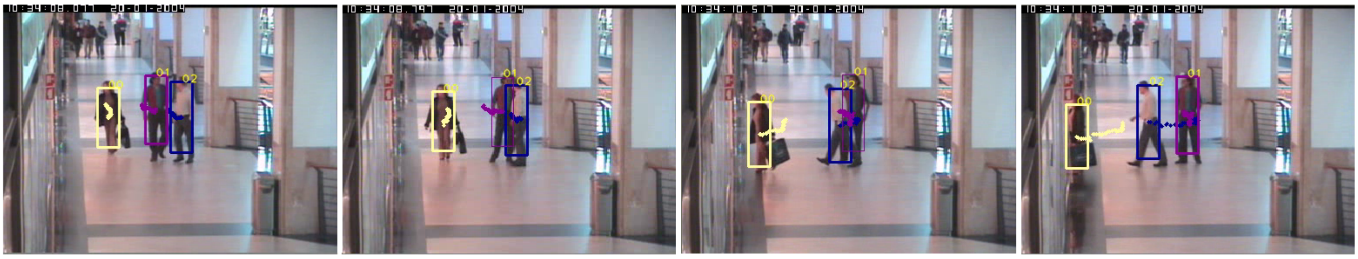


Figure 5. Results for the Caviar dataset. The occluded object is depicted with a thinner bounding box.

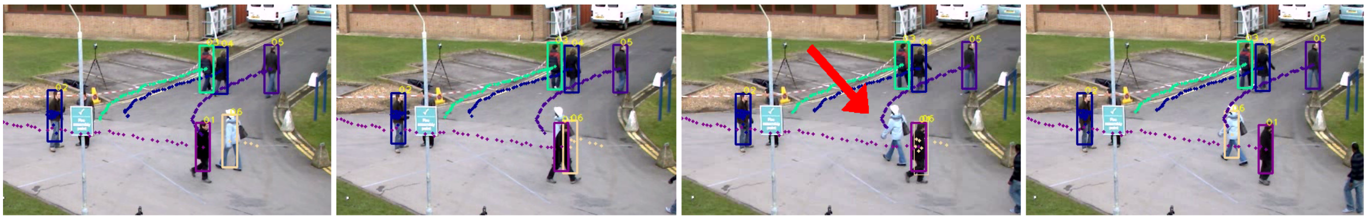


Figure 6. Results for the PETS 2009 dataset. Target ”06” regains its identity after the occlusion despite the template drifting.

[21] K. Zhang, L. Zhang, and M. Yang, “Real-time compressive tracking,” in *ECCV*, Springer, 2012.

[22] L. Matthews, T. Ishikawa, and S. Baker, “The template update problem,” *PAMI*, vol. 26, no. 6, pp. 810–815, 2004.

[23] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, “Image segmentation with a bounding box prior,” in *Computer Vision, 12th International Conference on*, IEEE, 2009.

[24] C. Rother, V. Kolmogorov, and A. Blake, “Grabcut: Interactive foreground extraction using iterated graph cuts,” in *ACM Transactions on Graphics (TOG)*, vol. 23, pp. 309–314, ACM, 2004.

[25] S. Vicente, V. Kolmogorov, and C. Rother, “Graph cut based image segmentation with connectivity priors,” in *CVPR 2008*, IEEE, 2008.

[26] B. Yang, C. Huang, and R. Nevatia, “Segmentation of objects in a detection window by nonparametric inhomogeneous crfs,” *Computer Vision and Image Understanding*, vol. 115, no. 11, pp. 1473–1482, 2011.

[27] P. Forssen, “Maximally stable colour regions for recognition and matching,” in *CVPR*, IEEE, 2007.

[28] B. Wu and R. Nevatia, “Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors,” *International Journal of Computer Vision*, vol. 75, no. 2, pp. 247–266, 2007.