Detection and Tracking of Objects with Direct Integration of Perception and Expectation

Kai Jüngling, Michael Arens
Research Institute for Optronics and Pattern Recognition
Ettlingen, Germany
{juengling, arens}@fom.fgan.de

Abstract

One of the main challenges in video-based multi-target tracking is the consistent maintenance of object identities over time. We present a novel approach to that challenge that integrates tracking and detection in a single process. We thereby inherently solve the identity problem and gain additional stability of the object detection performance. For that purpose, we extend a state-of-the-art local-feature based object detector by integrating expectations resulting from tracking directly into the detection procedure on the level of features. By that combination of newly gathered and expected local features we are able to directly integrate new data-evidence with object knowledge collected in the past without changing the detection approach itself. Since our tracking approach is solely based on local features, without employing other features like color or shape, it works independently of underlying video-data characteristics and preserves the general applicability independent of object-class specifics.

1. Introduction

Object tracking in videos has been subject to extensive research over the past decades. Many of the traditional tracking approaches are based on a foreground detection that distinguishes objects from a static background by a subtraction ([16]). A drawback of these systems is the disability to reliably distinguish different object classes. Recent advances in object detection ([17], [18], [10], [15], [7], [14]) encourage the use of trainable, class-specific object detectors for tracking tasks in complex environments.

In this paper, we address the problem of detecting and tracking multiple objects in real-world environments from a monocular camera. Although we focus on detecting and tracking people, our approach is independent of object specifics and thus generically applicable for tracking any object class.

Our object tracking is based on a state-of-the-art feature based object detector, introduced by Leibe *et al*. in [10].

The initial work of Leibe *et al*. has been extended by many authors that set up an object tracking process on the feature based object detector. Leibe *et al*. build a tracking based on this detector in [11] and [12]. They introduced a technique to treat tracking and detection of objects as a combined optimization problem and solved the assignment problem by an MDL (Minimum Description Length) approach. In course of this, they focused on the long-term track refinement and trajectory estimation with the standard detection approach.

Andriluka *et al*. introduced a method of combining tracking and detection of people in [3]. This approach uses knowledge of the walking cycle of a person to predict a persons position and control the detection. A prerequisite of this approach is the annotation of training data with body parts. Based on this information they integrate prior model knowledge on possible articulations and model the temporal coherency within a walking cycle of a person.

In [15], Seemann *et al*. extended the Implicit Shape Model (ISM) based object detector by the inclusion of knowledge of the specifics of articulated objects. In [14], they additionally introduced a technique to build object-instance models on-the-fly. These models are used to detect specific class instances in input images based on previously seen feature-configuration.

In contrast to all these former extension, we choose a different approach which automatically solves the assignment problem in tracking and needs no further assumptions on the objects to be tracked. We unite object tracking and detection in a single process and thereby address the tracking problem while enhancing the detection performance. The coupling of tracking and detection is carried out by a projection of expectations resulting from tracking into the detection on the feature level. This approach is suited to automatically combine new evidence resulting from sensor data with expectations gathered in the past. By that, we address the major problems that exist in tracking: we automatically

2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops
978-1-4244-4441-0/09/$25.00 ©2009 IEEE

preserve object identity by integrating the expectation into the detection, and, by using the normal codebook-matching procedure (see [10]), we automatically integrate new data-evidence into existing hypotheses. The projection of expectation thus stabilizes the detection itself and reduces the problem of multiple detections generated by a single real-world object. By adapting the weights of projected features over time, we automatically take the history and former reliability of a hypothesis into account and therefore get by without a special approach to assess the reliability of a tracked hypothesis.

In contrast to [11], our approach solely builds on local features used for detection and tracking without calculating any additional features like color or shape. As opposed to [3], we neither integrate specific model assumptions regarding the character of object classes nor the behavior of those. This would reduce the applicability of the tracking to specific object classes and could be counter-productive in situations where object behavior differs from the trained model. By that, our approach preserves general applicability and works completely independent of the characteristics of the underlying video-data and especially of object-class specifics.

The remainder of this paper is structured as follows: Section 2 briefly introduces the object detector and the extensions we applied. Section 3 points out the main contribution of this paper, the feature based integration of tracking and detection. Section 4 shows the improvements that our approach gains compared to separated tracking and detection and presents a quantitative evaluation of our tracking in different image sequences.

2. The refined object detection approach

In this section, we briefly describe the training and detection approach and the enhancements we made to the object detector described in [10].

2.1. Training

In the training stage, a specific object class is trained on the basis of annotated example images of the desired object category. The training is based on local features that are employed to build an appearance codebook of a specific object category. Leibe *et al*. use a combination of multiple cues to find interest points in the image and then use local Shape Context Descriptors [5] as feature description. Since this combination of multiple interesting point detectors is very time consuming, we use the SURF (Speeded Up Robust Features) descriptors described in [4]. This combination of feature point detection and feature description is specially designed for fast calculation. The features extracted from the training images on multiple scales are used to build an object category model. For that purpose, features are first clustered in descriptor space to identify reoccurring features that are characteristic for the specific object class. To generalize from the single feature appearance and build a generic, representative object class model, the clusters are represented by the cluster center (in descriptor space). At this point, clusters with too few contributing features are removed from the model since these cannot be expected to be representative for the object category. The feature clusters are the basis for the generation of the ISM that describes the spatial configuration of features relative to the object center and is used to vote for object center locations in the detection process.

2.2. Detection

To detect objects of the trained class in images, SURF features are extracted in input images. These features (the descriptors) are then matched with the codebook, where codebook entries with a distance below a threshold t_{sim} are activated and cast votes for object center locations. To allow for fast identification of promising object hypothesis locations, the voting space is divided into a discrete grid in x-, y-, and scale-dimension. Each grid that defines a voting maximum in a local neighborhood is taken to the next step, where voting maxima are refined by mean shift to accurately identify object center locations. At this point we make two extensions to the work of Leibe *et al*.

First, we do not distribute the vote weights equally over all features and codebook entries, but use the feature similarities to determine the assignment probabilities. By that, features more similar to codebook entries have more influence in object center voting. The probability $p(C_i|f_k)$ for an assignment of an image feature f_k and a codebook entry C_i is determined by:

$$p(C_i|f_k) = \frac{\rho(f_k, C_i) + t_{sim}}{t_{sim}}. \quad (1)$$

Where $\rho(f_k, C_i)$ is the euclidean distance in descriptor space multiplied by -1. The same distance measure is used for the probability $p(V_{\vec{x}}|C_i)$ of a vote for an object center location $\vec{x}$ when considering a codebook entry C_i. The vote location $\vec{x}$ is determined by the ISM that was learned in training. Here, $\rho(f_k, C_i)$ is the similarity between a codebook representative and a training feature that contributes to the codebook entry.

The overall probability for and weight of a vote $V_{\vec{x}}$ is:

$$V^w_{\vec{x}} = p(C_i|f_k) \cdot p(V_{\vec{x}}|C_i). \quad (2)$$

Second, we address the problem of the training data dependency. The initial approach by Leibe *et al*. uses all votes that contributed to a maximum to score a hypothesis. As a result, the voting and thus the hypothesis strength depends on the amount and the character of training data. Features,

which have often been seen in the training data result in codebook entries with a large amount of contributing features and thus in a vast of votes for a single object center location with only the evidence of a single image feature. This can result in false positive hypotheses with a high strength, generated by just a single or very few false matching image features. To solve this issue, we only count a single vote – the one with the highest probability – for a single image-feature. We hold this approach to be more plausible since a single image feature can only provide evidence for an object hypothesis once. As a result of this, overlaps in the hypotheses feature sets are automatically eliminated which leads to disjoint hypotheses. Again, we hold this to be more plausible since a single image feature cannot provide evidence for multiple objects.

The score δ of a hypothesis γ can thus directly be inferred by the sum of all I contributing votes:

$$\delta\gamma = \sum_{i=1}^{I} V_i^w. \tag{3}$$

Certainly, this score is furthermore divided by the volume of the scale-adaptive search kernel (see [10] for details), which is necessary because objects at higher scales can be expected to generate much more features than those on lower scales.

The result of the detection step is a set of object hypotheses Γ, each annotated with a score γ_ϕ. This score is subject to a further threshold application. All object hypotheses below that threshold are removed from the detection set Γ.

3. Feature based integration of tracking and detection

The object detection approach described up to now works exclusively data-driven by extracting features bottom-up from input images. At this point, we introduce a tracking technique that integrates expectations into this data-driven approach. The starting point of the tracking are the results of the object detector applied to the first image of an image sequence. These initial object hypotheses build the basis for the object-tracking in future. Each of these hypotheses consists of a set of image features which generated the according detection. These features are employed to realize a feature based object-tracking.

3.1. Projection of object hypotheses

For every new image of the image sequence, all hypotheses Γ known in the system at this time T, each comprising a set of features Π_γ, are fed back to the object detection before executing the detection procedure.

For the input image, the SURF-Feature-Extraction is performed, resulting in a set of image features Φ^{img}. The object detection procedure, described in section 2.2 is now executed for each known object hypothesis independently, whereby the feature set Π_γ is projected into the image. For that, we predict the feature's image positions for the current point in time (a Kalman-Filter that models the object-center dynamics assuming constant object acceleration is used to determine position prediction for features. Note that this is thought to be a weak assumption on object dynamics) and subjoin these feature to the image features.

In this joining, three different feature types are generated: The first feature type, the *native image features* Φ^{img} refers to features that are directly extracted from the input image. These features contribute with the weight $P_{type=nat}$, which is set to 1.

The second feature type, the *native hypothesis features*, is generated by projecting the hypothesis features Π_γ to the image. These features are weighted with $P_{type=hyp}$ and are added to the detection-feature-set Π_γ^{tot} of hypothesis γ:

$$\Pi_\gamma^{tot} = \Phi^{img} \cup \Pi_\gamma. \tag{4}$$

These features integrate the expectation into the detection and the weight is set to a value in the range $[0-1[$.

The next step generates the features of the third type, the *hypothesis features with image feature correspondence*. For this purpose, the hypothesis features Π_γ are matched (similarity is determined by an euclidean distance measure) with the image features Φ^{img}. Since (i) the assignment of hypothesis to image features includes dependencies between assignments and since (ii) a single hypothesis feature can only be assigned to one image feature (and vice versa), a simple "best match" assignment is not applicable. We thus solve the assignment problem by the the revised Hungarian method presented by J. Munkres in [13]. By that the best overall matching assignment and mutual exclusivity is ensured.

Feature assignments with a distance (in descriptor space) exceeding an assignment threshold κ_{feat} are prohibited. An additional image-distance constraint for feature pairs ensures the spatial consistency of features. Every $\phi \in \Phi_{img}$ which has a $\pi \in \Pi_\gamma$ assigned, is reassigned to feature type 3 and contributes with the weight $P_{type=mat}$ (the matching hypothesis feature π is removed from the detection set: $\Pi_\gamma^{tot} = \Pi_\gamma^{tot} \setminus \pi$ to not count features twice). This weight is set to a value > 1, because this feature type indicates conformity of expectation and data and thus contributes with the highest strength in the voting procedure.

The feature-type-weight is integrated into the voting by extending the vote-weight (see equation 2) with factor P_{type} to

$$V_{\vec{x}}^w = p(C_i|f_k) \cdot p(V_{\vec{x}}|C_i) \cdot P_{type}. \tag{5}$$

The voting procedure – which is the essential point in object detection – is thus extended by integrating the three different feature types that contribute with different strengths.

3.2. Coupled tracking and detection

From now on, the detection is executed following the general scheme described in section 2, split for each existing hypothesis independently using the feature set Π_γ^{tot} which now potentially contains features of all three types and thus joins expectation and data.

In every independent execution of the detection procedure, only the detection most likely to be the successor of the projected hypothesis is retained. This is the highest-scored detection which contains features of the second or third type. By that approach, the assignment of the new detections to existing hypotheses is conclusive and identity preservation therefore is inherently included in the procedure.

To detect new objects in the input image, a separate run is carried out on a reduced feature set. All image features, that have already contributed to a detection (that has been assigned to a hypothesis) in a former step, are removed from the set of native image features

$$\Phi^{img} = \Phi^{img} \setminus \{\Gamma_1 \cup \Gamma_2 \cup .. \cup \Gamma_N\}. \tag{6}$$

Where Γ_i is the set of image features, that voted for the i-th detection. This practice ensures, that features that have already contributed do not generate detections twice, but new objects are still detected independently of existing hypotheses.

The inclusion of factor P_{type} into the vote weight (5) has the result that image features are more likely assigned to a hypothesis with a feature match than to a hypothesis without a matching feature. This is important in cases where multiple objects overlap and features thus can often not be assigned to a hypothesis unambiguously. In these cases, the hypothesis, which had this feature assigned most often in the past, wins (this will be pointed out in section 3.4).

The novel detection approach is shown in figure 1 for an exemplary hypothesis. Image (a) visualizes the native features extracted from the input image. The projection of the hypothesis features and the different feature types are shown in image (b). We see that all 3 types of features contribute to the detection that is supported by the projected hypothesis. New image features (white), that vote for the same object center, are automatically integrated into the hypothesis. Purely projected hypothesis features (red) integrate expectations and support the detection preservation. Projected features with matching image features (blue) are increased in their significance for the modeled entity since they got repeated support by image data. Image (c) shows the last detection step after all features, which already have contributed to a detection, have been removed. One can see that only objects which have not been seen previously remain strong enough to generate a new detection.

3.3. Inherent reliability adaption

A detection resulting from a projected hypothesis already contains the correctly updated information since the integration of old and new information has been conducted in the detection process itself. The reliability of this detection thus already reflects the hypothesis reliability with inclusion of the hypothesis history. This is due to the inclusion of projected features into the detection. To achieve automatic adaption of the reliability over time, we replace the constant factor P_{type} in equation 5 by a feature-specific, time-varying function $P_{type}^{\pi,t}$ which is adapted every time the feature contributes to a detection. By that, feature history is inherently included. To accomplish this, the type factor $P_{type}^{\pi,t}$, of a feature $\pi \in \Pi_\gamma$ is set to

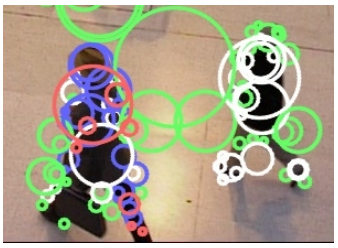
$$P_{type}^{\pi,t} = P_{type}^{\pi,t-1} \cdot \alpha_{type} \tag{7}$$

at time t. $P_{type}^{\pi,t-1}$ is the previous type-weight and α_{type} is the type-determined adaption rate previously used for the constant case (we use $\alpha_{nat} = 1$, $\alpha_{hyp} = 0.9$, and $\alpha_{mat} = 1.1$).

This rule leads to an automatic adaption of the feature weight determined by the presence of data-evidence. Initially, all features have the same type weight 1 since they have all been generated from native image features the first time they have been perceived. Afterwards, the adaption depends on whether there is new data-evidence for this feature or not. If a feature has an image match and is included in the according detection, its weight is increased because α_{mat} is > 1. Features that are permanently approved by data therefore are increased steadily over time. This leads to an automatic increase in hypothesis reliability that is determined by the weight of the assigned features.

When projected features are not supported by image features, the weight P_{type}^{π} is decreased because the factor α_{hyp} is < 1. The reliability of a hypothesis thereby automatically decreases when no new feature-evidence is available. In this case, the hypothesis is maintained by just the projected features. This inherent preservation of hypotheses even when no evidence is available, is essential to be able to track objects that are completely occluded for a short period of time. The period of time that a hypothesis is maintained in cases where no or very little image evidence is available, depends on the value of α_{hyp}. The lower this value, the faster hypothesis reliability decreases. Since these projected features are fed into the detection at every point in time, the hypotheses automatically re-strengthens when these features can be re-validated by image data after the occlusion occurred. New image features that are integrated into the detection (by voting for the same center location) also increase the reliability since they provide additional feature support for the hypothesis.

(a) *Native SURF features extracted from the input image.*



(b) *Projected hypothesis prediction after matching.*

(c) *Features used in the last detection run, after previously used features have been removed.*

Figure 1. Steps in the enhanced detection procedure. Different feature types are visualized by different colors. Green: Image features not contributing to any detection. White: Image features contributing to a detection. Red: Projected hypothesis features without a matching image feature. Blue: Projected hypothesis features with a matching image feature.

3.4. Automatic generation of object-identity models

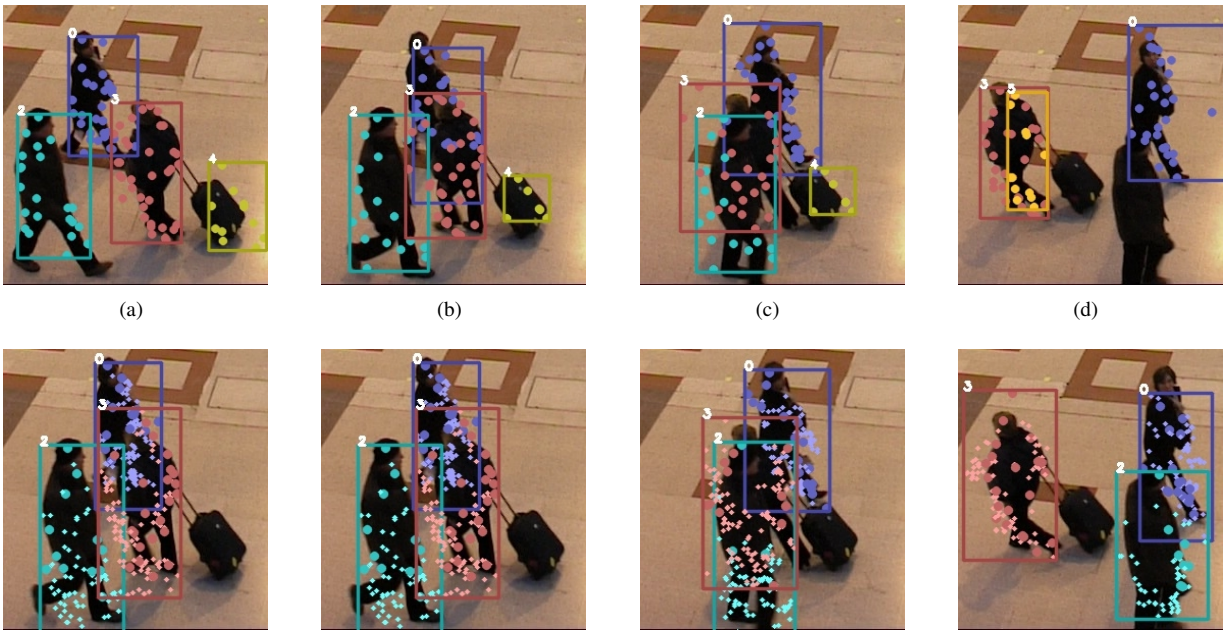
Beside the automatic adaption of reliability in the object detection step, the inherent inclusion of the feature history results in a second advantage. By this, features that have been seen very often and thereby have a high P^{π}_{type}, also have a strong vote in the detection process. Features that have not been seen in recent history, decrease in their influence in the object detection and are removed completely after a certain time (by a threshold applied to P^{π}_{type}). This is important in cases where the visual appearance of an object changes due to viewpoint changes or variances in lighting conditions. Features that are not significant for the object any more are removed after a certain time of absence. New features which are significant for the object now, are integrated into the hypothesis automatically. By this inherent generation of an *object-identity-model*, we are able to reliably re-identify objects based on the standard feature codebook without the need to establish an instance-specific codebook like proposed in [14]. Thus, we keep the generality of object description and simultaneously are able to re-identify single object instances. The identity models are relevant especially in cases where multiple objects occlude each other. Without the projection of hypotheses, this situation often results in indeterminable voting behavior. In practice, the strongest voting maxima is often right between the objects, since this position in the voting space gets support by features of two existing objects. In our approach, this problem is solved by the expectation projection and especially through the adaption of weights which generates the distinguishable object-identity model. By matching hypothesis- with image-features before detection and consecutively adapting the weight of the resulting votes by inherently including the feature history, we can determine which image features belong to which hypotheses. Features which have been seen in a hypotheses very often are, by a high P^{π}_{type}, more likely to be assigned to this hypothesis (see 3.1).

4. Results and evaluation

To assess the quality of our tracking compared to a feature-based tracking without the projection of expectations, we consider a situation (see figure 2) where persons should be tracked over significant occlusion. The top row shows results of a feature based tracking with independent detection and subsequent track formation. Here, we see that the features are interchanged between the occluding persons and a single person generates multiple object hypotheses in (d). As a result, hypothesis 2 is not correctly maintained since the features generated from the according person voted for another hypothesis.

The bottom row shows the results of our tracking approach in the same situation. As we see, the object-identities are preserved correctly and only very few features are permuted between the objects, although three object overlap significantly in this situation. This is the prerequisite for the ability of identity preservation in such difficult situations, that is, as we see, correctly maintained. Additionally, generation of multiple hypotheses on a single object is prevented by the stabilizing effect of feature projection. False positives that occur on the luggage in the top-row are prevented by projection too, because the threshold for hypothesis generation can be set stricter (thanks to the additional support of projected features).

For quantitative evaluation, we chose a subset of sequences from the PETS 2006 (Performance Evaluation of Tracking Systems) dataset [2]. Here, a scene is observed by a static camera with a resolution of 720x576 and people are visible side-face. We chose this data for evaluation, because it comprises different situations where especially identity preservation is difficult to accomplish. We picked four (hereafter referred to a sequence 1-4) image sequences that comprise different difficulties for a tracking system, to specifically assess the performance of our approach in different situations. To show the applicability of our tracking approach in a typical surveillance scenario, we addi-



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2. Comparison of tracking results using the integration of perception and expectation (bottom-row) and a feature-based tracking without these extensions (top-row). Colorized dots visualize the features contributing to a person-hypothesis. Small dots in the bottom-row depict projected features that can not be verified by image data but contribute to the according hypothesis.

tionally evaluated the tracking in a single sequence (hereafter referred to as sequence 5) of the challenging CAVIAR Test Case Scenarios [1] where people are observed with a wide angle camera with a resolution of 384x288. Here – as shown in figure 3 – people appear very small and with a low contrast to background.

We annotate every person in the video sequence with a bounding box. Since our tracking approach is in principle capable to infer the presence of occluded objects when they have seen before (see figure 2), we also annotate temporary occluded persons and the occluded parts of persons if they have been "fully-visible" previously in the sequence.

To determine whether an object hypothesis is a true- or a false positive, we use the *overlapping* criterion. This assesses object hypotheses using the ground-truth and hypotheses bounding boxes. The overlap between those is calculated by the Jaccard-Index [9] (compare intersection-over-union criterion [8]). Only a single hypothesis is counted per ground-truth object, all other hypotheses are counted as false positive for this object. We demand a minimum overlap of 0.5 (50%) to be regarded as true positive.

To assess tracking performance, we use the metrics:

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_{i,t} gt_t^i} \tag{8}$$

$$MOTA = 1 - (\overline{m} + \overline{fp} + \overline{mm}) \tag{9}$$

from [6], which include a complete assessment of tracking performance, detection performance and precision.

The Multiple object tracking precision (MOTP) indicates the overall exactness of detections where d_t^i is the distance (i.e. the center distance) of a true positive detection and the ground truth. Since we evaluate our tracking performance using a bounding box criterion, we do not use the distance but the overlap of detection and ground-truth bounding box. Thus, MOTP in our case is the mean bounding box overlap (though 1.0 would be the best results here) of all correct detections.

The multiple object tracking accuracy (MOTA) accounts for the overall tracking performance by taking into account the miss-ratio $\overline{m}$, the false positive ratio $\overline{fp}$ and the mismatch ratio $\overline{mm}$:

$$\overline{m} = \frac{\|m\|}{\|gt\|} \;,\; \overline{fp} = \frac{\|fp\|}{\|gt\|} \;,\; \overline{mm} = \frac{\|mm\|}{\|gt\|} \tag{10}$$

that are accumulated over all frames ($\|gt\|$ is the number of objects in ground truth). To allow for comparison of our results with work like [11], [12], or [3], we additionally show the recall (ratio of true positives and ground-truth objects) and the false positives per image in the result table 1.

The following evaluation is carried out with a person detector trained on 59 person appearances. For sequence 5, we trained a separate detector on 27 person appearances. The

	Frames	Objects (#ids)	MOTP	Miss rate	False pos. rate	Mismatches	**MOTA**	Recall	False pos./image
Seq1	381	487 (4)	0.64	0.14	0.09	0	**0.77**	0.86	0.11
Seq2	185	793 (10)	0.71	0.17	0.08	0	**0.75**	0.83	0.39
Seq3	250	808 (6)	0.61	0.27	0.24	0.01 (11)	**0.48**	0.73	0.86
Seq4	673	1607 (11)	0.65	0.12	0.12	0.002 (4)	**0.76**	0.88	0.29
Seq5	1043	2142 (3)	0.55	0.5	0.1	0.001 (3)	**0.4**	0.5	0.21

Table 1. Tracking results.

separate detector is necessary here due to the completely changed camera perspective.

The first evaluation sequence comprises 487 appearances of 4 different persons in 381 frames. Here, people mainly appear without any occlusions or overlapping between persons. In this rather simple sequence, we accomplish a MOTA of 0.77 (see table 1) where misses and false positives are nearly equal and both rather low. As was expected here, no mismatch occurred.

The second sequence with 793 appearances of 10 different persons in 185 frames, observes people moving in a queue with partial overlapping and a single person crossing the path of the other persons, which leads to occlusion of the people in the back. Detection examples for this sequence are shown in figure 4(a)-(c). Here, we accomplish ratios nearly equal to sequence one, which shows that our tracking deals very well with short-time occlusion.

The third sequence with a total of 808 appearances of 6 different persons in 250 frames, observes a group of 4 people entering and moving in the scene jointly whereby strong overlapping between people occurs. The main challenge here is to detect the individuals in the group and to correctly maintain the identities over time. This is very challenging here, because person appearances are very similar, which results in a large homogeneous region during overlapping of persons. Here, our tracking accomplishes a MOTA of 0.48 which is inconceivable with time-independent detection and subsequent assignment of detections to tracks. Detection examples for this sequence are shown in figure 4(d)-(f). (Note that we did not evaluate the persons in front of the train because this would distort the tracking results for the specific group-situation)

The fourth sequence, with a total of 1607 appearances of 11 different persons in 673 frames is a typical tracking sequence where single persons move around, people overlap and moving paths cross. Here, we gain an overall MOTA of 0.76 with only 4 mismatches which shows the good performance of our tracker in general tracking situations.

Sequence five with a total of 2142 appearances of 3 person in 1043 frames evaluates the tracker performance in a sequence of the CAVIAR Test Case Scenarios. The main challenge here is to track persons appearing very small with a very low contrast to the background (see figure 3). The results in table 1 show the good performance with a recall

Figure 3. Sample image of sequence 5. Persons are marked with red boxes.

of 0.5 at 0.21 false positives per image even in this difficult scenario. The unusual high miss-rate results from people vanishing in the back of the scenario and from people that are difficult to detect in an area with glaring light.

5. Conclusion

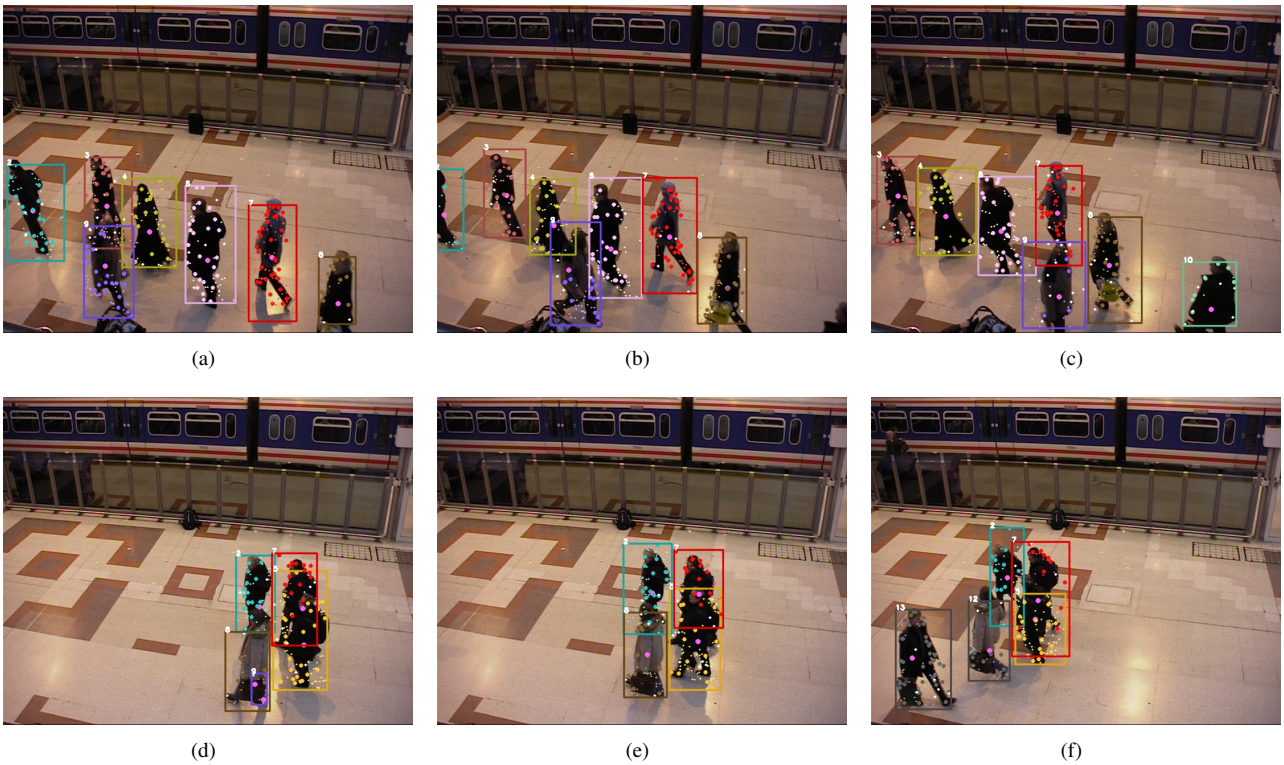
In this paper, we presented a novel feature-based object tracking strategy that unites tracking and detection by joining expectations and data on the level of local features. We evaluated our tracking in situations that comprise different difficulties for person tracking and showed that, additionally to the good performance in person tracking in general, our approach is especially able to keep person identity across occlusions. This is not done by heuristics that capture special situations, but inherently included in our approach. In these situations, we are – by using the feature projection – furthermore able to estimate the person by a bounding box, even when this person is fully occluded.

References

[1] Ec funded caviar project, url: http://homepages.inf.ed.ac.uk/rbf/caviar/. 2001.

[2] In *PETS 2006. 9th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, New York, USA, June 2006. (see `http://www.cvg.rdg.ac.uk/PETS2006/index.html`).

[3] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Proc. IEEE*



(a) (b) (c)

(d) (e) (f)

Figure 4. Example detections from evaluation sequences two and three. Sequence 2:(a)-(c), Sequence 3:(d)-(f).

Conference on Computer Vision and Pattern Recognition (in Press), Anchorage, USA, June 2008.

[4] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *Proc. 9th European Conference on Computer Vision*, pages 404–417, Graz, Austria, May 2006.

[5] S. Belongie, J. Malik, and J. Puchiza. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[6] K. Bernardi, A. Elbs, and R. Stiefelhagen. Multiple object tracking performance metrics and evaluation in a smart room environment. In *The Sixth IEEE International Workshop on Visual Surveillance*, Graz, Austria, May 2006.

[7] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*, 2006.

[8] M. Everingham et al. The 2005 pascal visual object class challenge. In *In Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, 2006.

[9] P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise de Sciences Naturelles*, 4(3):223–370, 1908.

[10] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77:259–289, 2008.

[11] B. Leibe, K. Schindler, N. Cornelis, and L. V. Gool. Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(10):1683–1698, 2008.

[12] B. Leibe, K. Schindler, and L. V. Gool. Coupled detection and trajectory estimation for multi-object tracking. In *Proc. International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brasil, October 2007.

[13] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society of Industrial and Applied Mathematics*, pages 32–38, 1957.

[14] E. Seemann, M. Fritz, and B. Schiele. Towards robust pedestrian detection in crowded image sequences. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Mineapolis, USA, June 2007.

[15] E. Seemann, B. Leibe, and B. Schiele. Multi-aspect detection of articulated objects. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1582–1588, New York, USA, June 2006.

[16] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 246–252, Ft. Collins, CO, USA, June 1999.

[17] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 511–518, Kauai, HI, USA, December 2001.

[18] 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*, 75(2):247–266, 2007.