# Hierarchical object detection and tracking with an Implicit Shape Model

**K. Jüngling**[1], **S. Becker**[1], **and M. Arens**[1]
[1]Object Recognition, Fraunhofer IOSB, Ettlingen, Germany

**Abstract**— *One important field in machine vision is object tracking. In most real-world applications, multiple object instances of different classes which influence each others appearance are of interest. Independent treatment of these objects in detection and tracking is not sufficient. In this paper, we present an object tracking algorithm which takes this into account and builds on an Implicit Shape Model (ISM) based detection and tracking. This suits it for the task of tracking multiple interacting objects in a cascaded approach because the local feature based ISM approach allows inference of the image regions classification was based on and hence allows for differentiation of the objects on the sensory level. We conduct two experiments: The first experiment demonstrates tracking on a surveillance dataset for the case of persons. In the second experiment, hierarchical tracking is performed for the cases of persons on ships and bags which are carried by persons.*

**Keywords:** Object detection, object tracking, implicit shape model, SIFT

## 1. Introduction

One important field in computer vision, which is relevant in many application areas is object detection and tracking. Over the years, many approaches have been proposed for these areas. With the progress made in the area of local image features in the past few years, approaches for object detection and tracking have evolved from simple foreground and motion detection algorithms [1], [2], [3] to more sophisticated algorithms approaching the problem in a different way (see [4] and [5] for an extensive survey). Rather than treating a detected foreground region as an object instance of a certain class, these approaches [6], [7], [8], [9], [10] employ machine learning techniques to train object class specific models which are used to detect object class instances in input imagery. Many tracking approaches like [11], [12], [13], [14] build on these dedicated object detectors to perform tracking in image sequences. Main advantages of these approaches over motion or foreground based approaches are that (i) different object classes can be distinguished unambiguously, (ii) they work despite camera motion and (iii) they are most widely stable against environmental conditions. Good performance of these approaches both for detection and tracking of multiple instances of a single object class has been shown in several papers [14], [7], [15].

In many cases, not only a single object class is of interest, but many different classes are relevant in certain applications. In some cases, it is sufficient to detect and track objects of different classes independently from each other since no inherent relationship between object classes exists (note that this does not mean that object instances of the classes may not interact on signal level, but only that no inherent relationship between classes which can be known priorly exists). In some cases, inherent relations between object classes exist. Here, using the context of one object to detect and track instances of the other object class can be useful or even necessary in some cases. Examples for such relations are diverse, specifically when considering persons as one involved object class. For instance if driving cars are tracked, we can always expect that there is at least one person on board driving the car. Another example for a relation are bags or other objects which are carried by and thus can specifically be searched for in the context of persons. The mentioned cases are 'part-of' relations that typically exists between two object classes. This means that we can use a detected object of the one class to guide detection of the second class. Other than a part-of relation in a single object class, e.g. a wheel is part of a car, these cannot be identified directly using spatial reasoning which employs structural knowledge of the object class.

In this paper, we tackle the use of dependencies between object classes to guide, improve and enhance object detection and tracking. For this, we introduce a hierarchical object detection and tracking strategy which is based on the Implicit Shape Model [8] and SIFT features [16] and by that extend the work of Jüngling et al. [17].

The paper is structured as follows. Section 2 introduces the implicit shape model based detection and tracking strategy and presents the extension to this approach for hierarchical detection and tracking. Section 3 presents experiments for single class person tracking and two cases of hierarchical tracking where persons are involved once as 'containing class' and once as 'contained class'. Section 4 concludes.

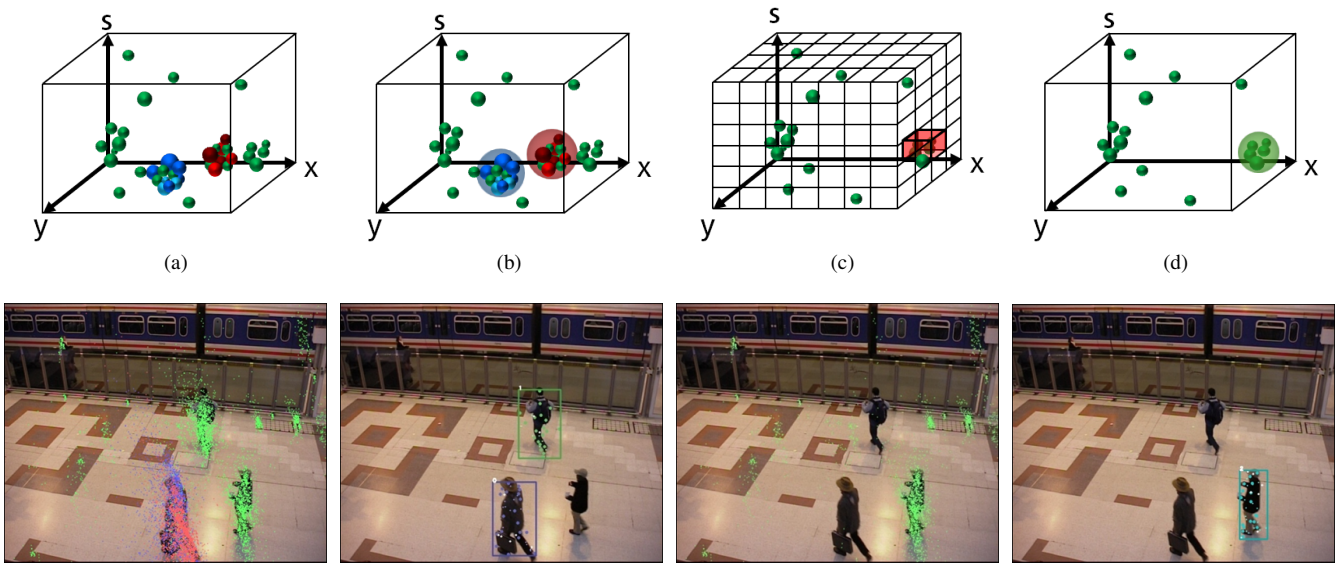


Fig. 1: Object tracking in the hough-voting space. (a) Joint voting space: Colors indicate vote type. Green votes do not have a correspondence to a former hypotheses feature and thus can vote for any object. Other colors have correspondences to former hypothesis features and vote only for specific objects hypotheses. (b) Mean-shift-search for tracking of known objects. Mean-shift search can be started for an existing hypothesis directly from the last known position of the hypothesis. No initial maxima search is necessary to identify possible object hypotheses since the starting points for the searches are already known. (c) Maxima search in the discretized hough-space to detect new objects which are not known in the system yet. (d) Mean-shift search for refinement of maxima position of new hypotheses.

## 2. Hierarchical object detection and tracking

The tracking approach this work builds on was introduced in [17], [14] for the case of persons. The key idea of this tracking is that the Implicit Shape Model (ISM), which is a trainable object detection approach that builds on local features (we use SIFT in this paper), is extended for tracking. The ISM object detector and thus the tracking works on the basis of a codebook for a specific object category which is built in a training step based on sample images of the object category of interest. Here, SIFT features found in the training samples are input to a clustering that identifies reoccurring features which are significant for the object category. The cluster centers are employed as prototypes for the codebook which describes the object category generically. Together with these prototypes, a spatial distribution, which encodes the position of features that contributed to this prototype relatively to the object center, is stored for each prototype. This codebook is used to detect persons in input images by matching SIFT features extracted from the input image to the prototypes. The spatial distribution of matching prototypes is then employed to build a hough-voting-space where each spatial distribution entry votes for a possible object center location. A mean-shift maxima search in this hough-space is conducted to detect persons. Tracking builds on this ISM detection and extends it by integrating temporal information into the hough-based object detection approach.

This temporal extension is performed on the level of SIFT features which describe the object hypotheses: All hypotheses already known in the system at a time T (this can be an empty hypotheses set – new objects are detected and integrated as hypotheses automatically) are predicted for the current instant of time on the level of features – we use a simple kalman filter model (this kalman filter could be replaced by other, more sophisticated dynamics modeling) to predict feature motion . The hypothesis features are then matched with SIFT features extracted from the current input image to build feature correspondences. These correspondences form the temporal information integration. This integrated information is then input to the object detection procedure, namely the features are matched with the appearance codebook to build the hough voting space. This voting space is visualized in figure 1 and has two major differences compared to the standard 'detection voting space' (see [17], [14] for details): (i) the vote weights are dependent on whether a feature correspondence could be formed or not, (ii) votes which result from features that had a correspondence to a former hypothesis feature are tagged with the specific hypothesis ID and can only vote for this certain hypothesis, all other votes can vote for any hypothesis. Figure 1 shows how tracking is performed in this

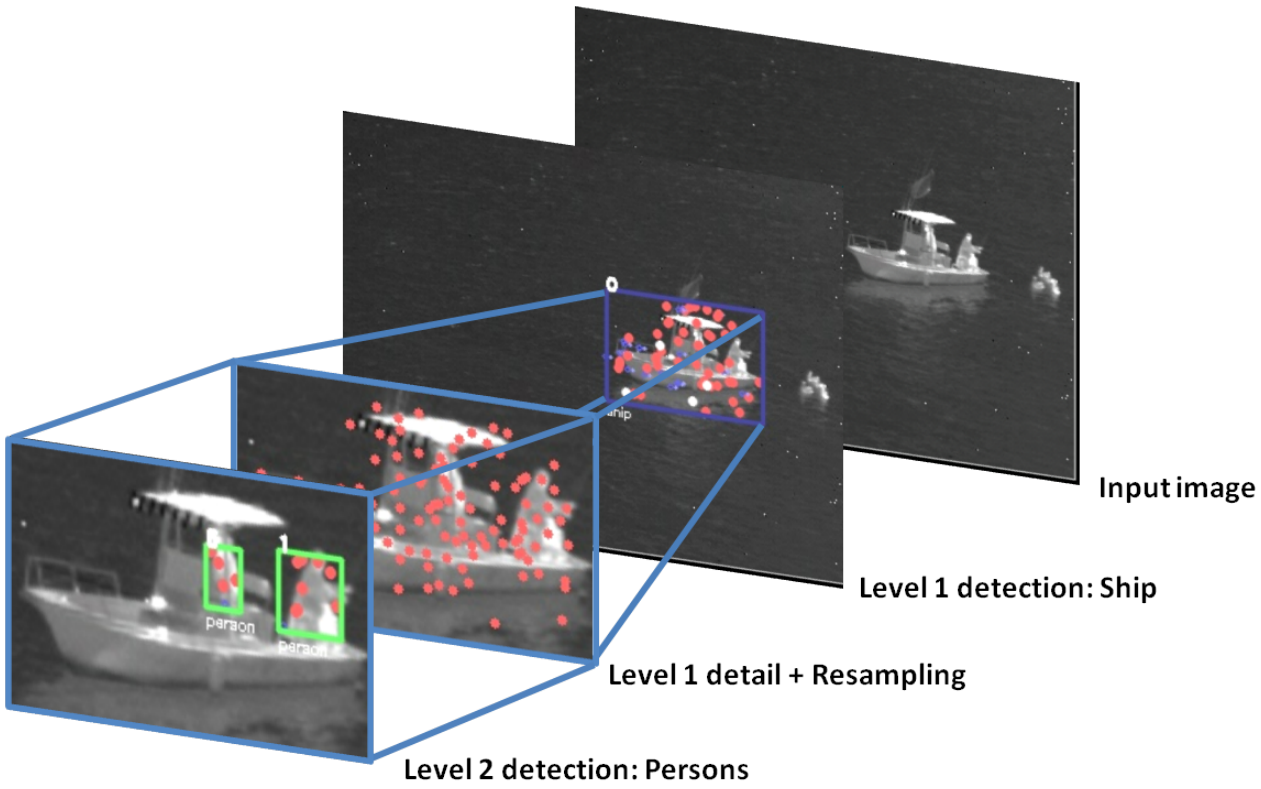


Fig. 2: Hierarchical object detection using the example of persons on a ship. First, ship detection is performed. SIFT features in the area of the ship bounding-box which contributed to the ship hypothesis are removed from further consideration. Remaining features are employed to perform person detection in the area of the ship. By this approach, false-alarms can be reduced and the range of detectable objects can be increased.

voting space: (a) shows the voting space. Here, green votes can vote for any object (also for new objects which have no hypothesis assigned at this instant of time), blue and red votes vote for the known object hypotheses with ID 1 and 2 respectively. As shown in figure 1 (b), tracking is performed by starting a mean-shift maximum search for every known hypothesis independently. This search only includes general votes (green) and votes for this specific hypothesis. After convergence of the mean-shift search, votes inside the mean-shift kernel are used to update the hypothesis. New objects which have not been seen before are detected in the 'reduced voting space' as shown in figure 1 (c) and (d). Here, all votes which already contributed to a hypothesis and those which vote only for a 'known object' are removed. All remaining votes can form new object hypotheses. For that, the standard object detection is performed. New object hypotheses are transfered to the set of 'known hypotheses'.

This tracking approach is suited to track multiple objects of the same class. It can also be employed to track objects of multiple classes. In this case, multiple codebooks can be build and employed to detect and track objects of different classes independently of each other. As motivated in the introduction, in some cases, this independent treatment of object classes might not be sufficient or at least not appropriate. In cases where inherent relationships between object classes or knowledge of a possible context in specific environments exists, this information can be employed to improve detection and tracking performance or in some cases even allow for detection and tracking which is not possible without that information. For instance, when considering a situation as in figure 2 where two person are on a ship. Direct detection of these person is very difficult since they are not visible very well and appear at a low resolution under short term occlusions. In this case, the knowledge about the fact that it is very likely that at least one person appears on the ship can be used to form a part-of relation between the ship and a person. Additionally, the knowledge that it is unlikely that persons appear in the water surrounding the ship can be used to confine the search space for persons and by focusing on the area determined by the ship detection reduce 'person false-alarms' which might occur on the water.

In case of this part-of relation between the objects (ships and persons), we can use the bounding box of the 'holding object' as starting point for detection of objects of the second class. This can be done using any possible object detector (also a motion detection) which provides us with

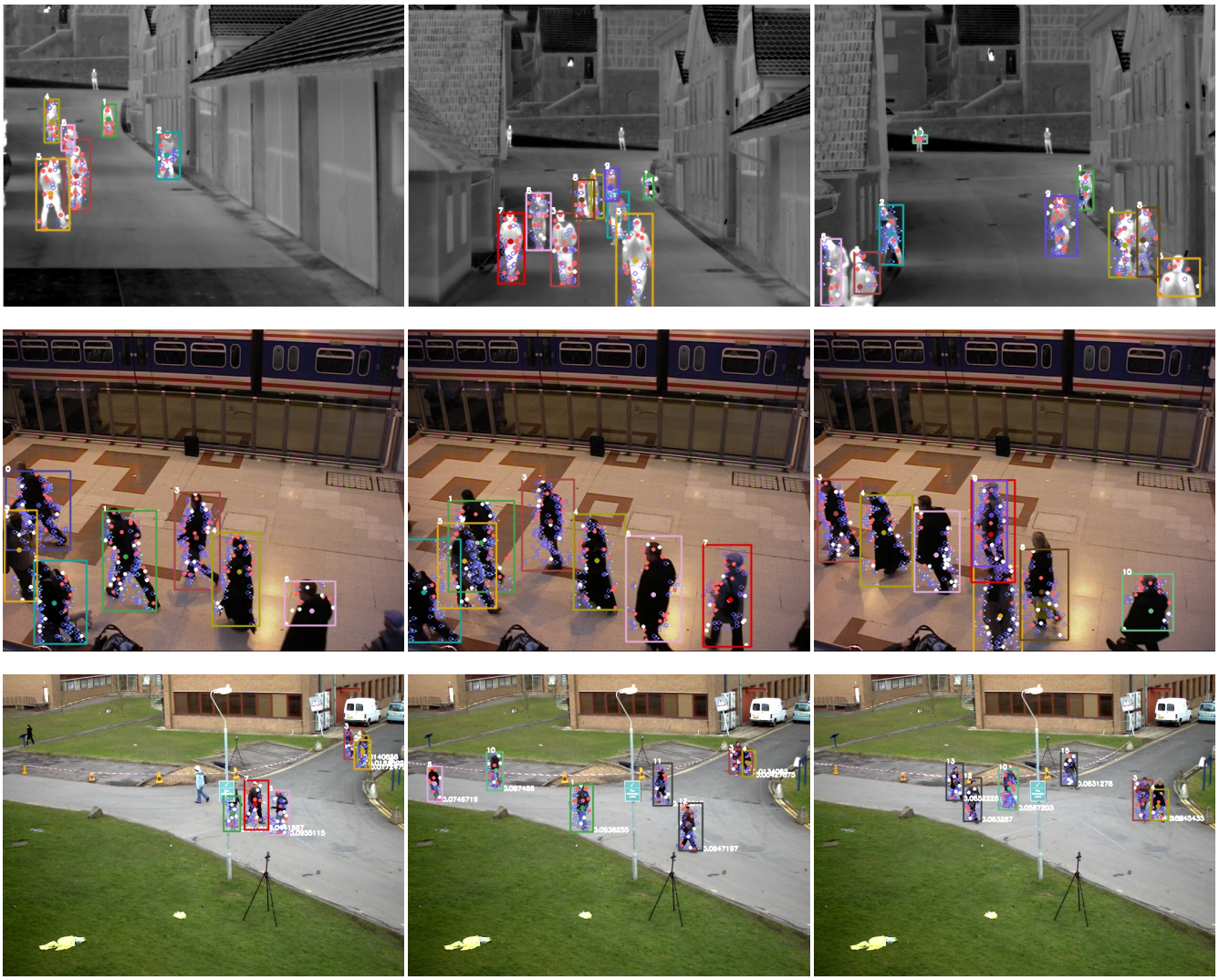

Fig. 3: Tracking results of different scenarios.

a bounding box. The major advantage of the ISM object detector is that not only the object surrounding bounding box is known, but also the SIFT image features which contributed to the 'surrounding object hypothesis' can be inferred. Consequently, the image parts which already contributed to the holding object are known. Since an single image region (more specifically a single pixel) cannot provide evidence for two objects, these regions can be excluded in the search of objects of the second class since we already know that these regions belong to the first object. This is not possible with motion or foreground segmentation and even not with other dedicated object detectors as the HOG detector or neural networks. In case of the ISM detector, we can do this inference based on the SIFT features since we know which features already contributed to the hypothesis of the surrounding object. To detect objects of the second class, we simply remove features which have already contributed – this is very similar to vote removal in tracking prior to detecting new objects – and afterwards perform detection using the set of remaining features in the area of the bounding box.

In some cases, subsumed objects are not visible very well or only visible at a low scale. Here, it is insufficient to perform object detection employing the set of remaining features in the bounding box area of the subsuming object. Additional preprocessing steps or different parametrization of feature extraction might be necessary. Such a case is shown in figure 2 where persons on a ship are to be tracked. As we see, persons are not visible very well due to low scale, partial occlusion and low contrast to the surrounding. To ensure stable detection and tracking of these persons, a more detailed inspection of the data is necessary compared to ship detection. This is facilitated by changing SIFT

Table 1: Tacking results in different scenarios.

| Sequence | Bonnland | Pets06 | Pets09 |
|---|---|---|---|
| Frames | 1664 | 185 | 795 |
| Objects (#ids) | 5426 (12) | 793 (8) | 4968 (20) |
| MOTP | 0.77 | 0.76 | 0.70 |
| $\overline{m}$ | 0.20 | 0.06 | 0.18 |
| $\overline{fp}$ | 0.12 | 0.008 | 0.09 |
| $\overline{mm}$ | 0.001 (3) | 0 | 0.004 (21) |
| **MOTA** | **0.68** | **0.93** | **0.73** |

parametrization to extract features (i) at lower scales (this is done internally by image size doubling) and (ii) with a lower contrast. This adds new features to the feature set which permit stable detection of person even under these difficult circumstances.

In addition to single image hierarchical object detection, objects are tracked over time. On level 1, tracking is performed as described before. On the following levels, tracking is performed employing the same principal, but using the set of remaining features of the preceding level. This tracking is performed in the reference frame of the preceding object using the bounding box as coordinate reference.

## 3. Experiments

To show the general applicability of the tracking approach, we evaluate single class tracking for the case of persons in different application scenarios shown in figure 3. As we see, tracking performs well in different scenarios ranging from infrared sequences acquired from a moving camera (first row) to typical surveillance scenarios (second and third row, dataset PETS06 and PETS09 workshops [18], [19]). The good results are confirmed by the evaluation shown in table 1 where standard CLEAR/MOT tracking performance assessment metrics MOTA (Multiple Object Tracking Accuracy) and MOTP (Multiple Object Tracking Precision) from [20] are used for evaluation. Note that person tracking performance is comparable to performance of current state-of-the-art person tracking approaches.

The technique of hierarchical object detection and tracking is specifically useful for two major purposes. First it can be used to detect objects which are not visible very well using contextual information. The second case is to confine the space where instances of certain object classes might occur. Such a case is shown in figure 4 where 'person' is the holding object class and the bag carried by the person is the object of interest in context. The left image shows detection results for the class 'bag' when the object detector is applied to the whole image without using contextual information. As we see, a lot of false alarms are generated throughout the image. While some of the false alarms might be eradicated by the use of a more narrow threshold for hypothesis score, some structures the visual appearance of

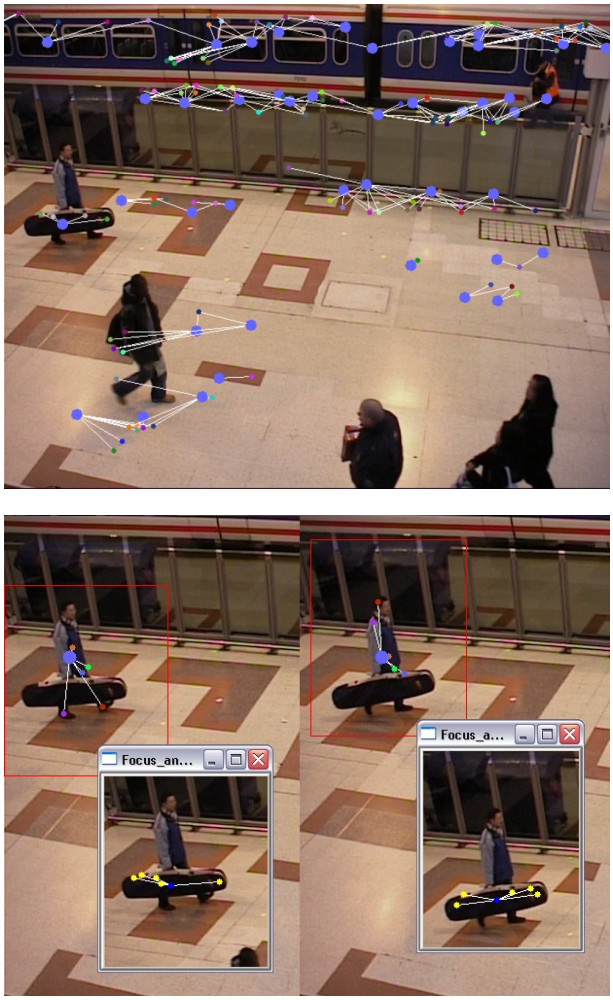


Fig. 4: Application of hierarchical object detection to the task of detecting bags in the context of a detected person. Left image shows bag detection results without using persons as context. As we see here, a lot of false alarms are generated by image structures which look like bags. The right image shows the result of using a detected person as context to confine the search space for bag detection. In this case, false alarm bag hypotheses on background structures can be removed.

which is very similar to the appearance of a bag – like in this case, horizontal structures on the train and the fence – will always generate false alarms. By using the context of the person to confine the search space (only bags which interact with a person are of interest), false alarms can be removed completely.

Although this approach for contextual object detection works well in these situations where both object classes are visible well enough to allow for identification as instance of a certain object class, there are some difficult situations, where the appearance of an object is very closely coupled

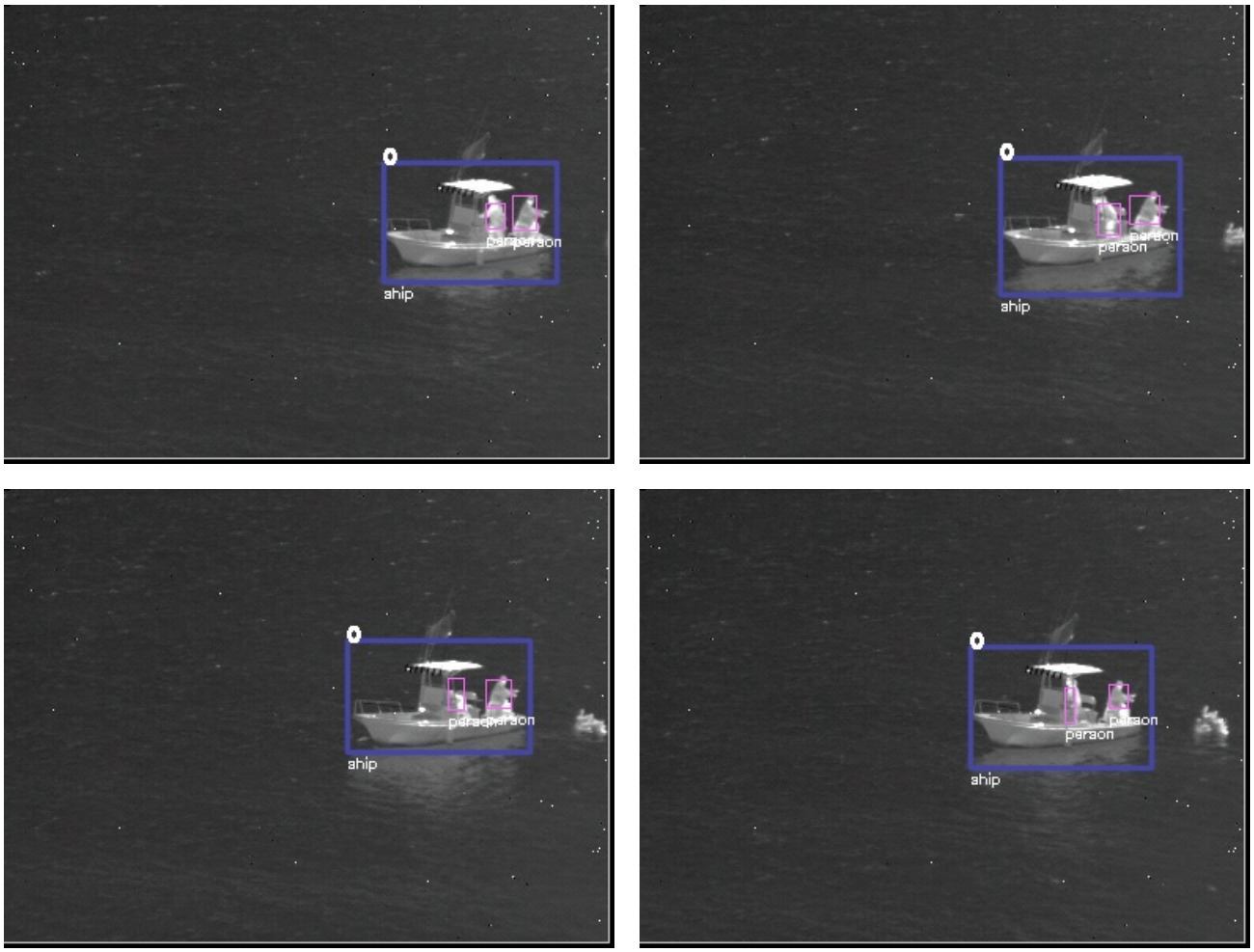


Fig. 5: Results of hierarchical tracking of persons on a ship.

to and influenced by the appearance of the 'holding object'. Examples for such cases are a backpack carried by a person or a person that is pointing a gun at someone. In these cases, the backpack and the gun are typically not visible well enough to identify them as instance of the respective class – even not when using hierarchical detection to confine search space. Here, more sophisticated approaches like [21] which treat detection of both classes as a coupled problem might be necessary. Tracking results of the first case relevant for hierarchical object detection and tracking are shown figure 5 and 6. Here persons are to be tracked on a ship. As we see, the persons themselves are not visible very well and a detection without the context of the ship would probably results in a lot of false alarms in the surrounding water and on the ship itself, without guarantee of correct detection of the person themselves. By employing hierarchical tracking of persons in the coordinate frame of the ship with removal of features that voted for the ship, persons can be tracked without any failure or false alarm.

# 4. Conclusion

In this paper we presented a SIFT and Implicit Shape Model based detection and tracking approach and its extension to a hierarchical detection and tracking approach which considers the context of object classes to (i) improve detection performance for other object classes and (ii) increase the range of situations and environments where objects of certain classes can be detected. We performed experiments for the case of person tracking as an example of multi-target and single object class tracking to show the good tracking performance in difficult situations. In addition, we performed experiments for hierarchical object detection and tracking for the case of bags which are carried by persons and persons which are present on a ship. For future work, more different object classes should be considered in experiments to show that hierarchical detection and tracking can be useful in many applications.

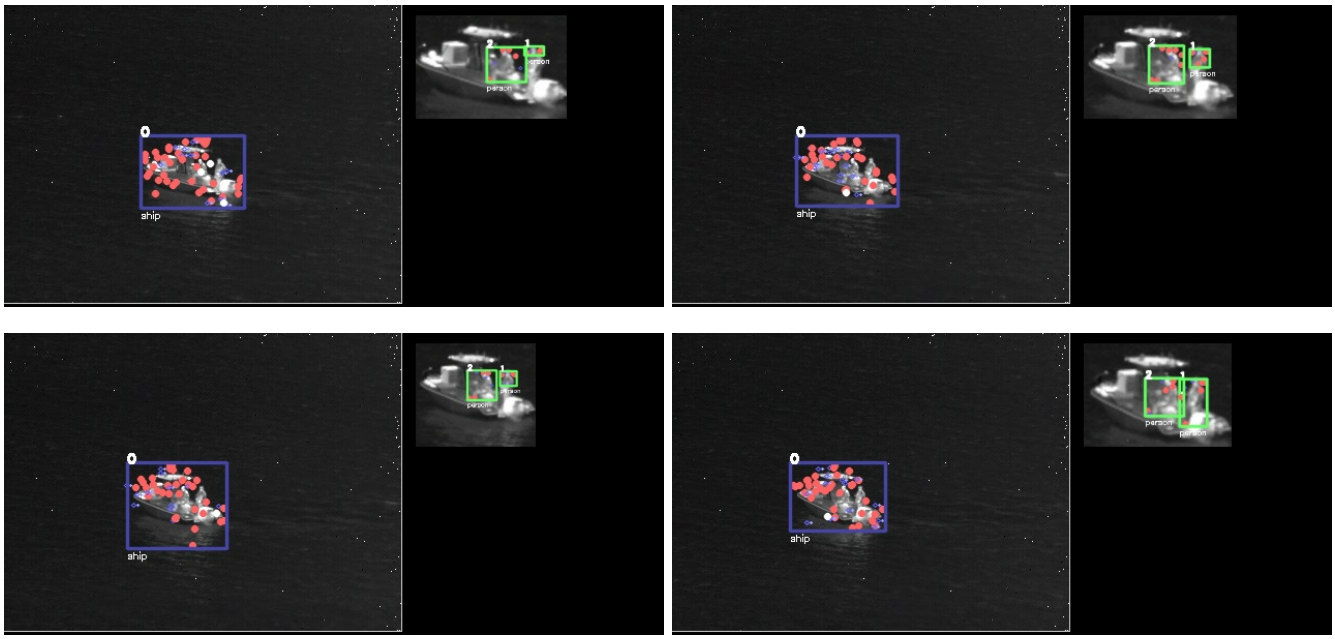


Fig. 6: Results of hierarchical tracking of persons on a ship. Left side shows ship tracking results. Right side shows the tracked ship and results of person tracking which was performed in the coordinate frame of the ship.

# References


[1] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1999, pp. 246–252.

[2] I. Haritaoglu, D. Harwood, and L. Davis, "W4s: A real-time system for detecting and tracking people in 2.5 d," in *Proc. European Conference on Computer Vision*, 1998, pp. 877–886.

[3] S. A. Berrabah, G. De Cubber, V. Enescu, and H. Sahli, "Mrf-based foreground detection in image sequences from a moving camera," in *Proc. International Conference on Image Processing*, 2006, pp. 1125–1128.

[4] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *Transactions on Pattern and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[5] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, vol. 34, no. 3, pp. 334–352, 2004.

[6] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[7] 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, November 2007.

[8] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 259–289, 2008.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Computer Vision and Pattern Recognition*, vol. 1, June 25–25, 2005, pp. 886–893.

[10] 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*, 2007, pp. 1–8.

[11] B. Leibe, K. Schindler, and L. V. Gool, "Coupled detection and trajectory estimation for multi-object tracking," in *Proc. International Conference on Computer Vision*, 2007, pp. 1–8.

[12] S. Gammeter, A. Ess, T. Jäggli, K. Schindler, B. Leibe, and L. Van Gool, "Articulated multi-body tracking under egomotion," in *Proc. European Conference on Computer Vision*, 2008, pp. 816–830.

[13] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *International Journal of Computer Vision*, vol. 73, no. 1, pp. 41–59, 2007.

[14] K. Jüngling and M. Arens, "Pedestrian tracking in infrared from moving vehicles," in *Intelligent Vehicles Symposium*, 2010, pp. 470–477.

[15] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. van Gool, "Online multi-person tracking-by-detection from a single, uncalibrated camera," *Transactions on Pattern Analysis and Machine Intelligence.*, 2010.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] K. Jüngling and M. Arens, "Detection and tracking of objects with direct integration of perception and expectation," in *Proc. Int. Conference on Computer Vision, ICCV Workshops*, 2009, pp. 1129–1136.

[18] "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).

[19] "Winter-pets 2009. 12th ieee international workshop on performance evaluation of tracking and surveillance," New York, USA, December 2009, (see http://winterpets09.net).

[20] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *Journal of Image Video Processing*, vol. 2008, pp. 1–10, 2008.

[21] S. Becker and K. Jüngling, "An implicit shape model based approach to identify armed persons," in *Proc. SPIE Defense, Security, and Sensing*, 2011, p. to appear.