Towards Robust Pedestrian Detection in Crowded Image Sequences

Edgar Seemann, Mario Fritz and Bernt Schiele
TU Darmstadt, Germany
{seemann,fritz,schiele}@mis.tu-darmstadt.de
http://www.mis.informatik.tu-darmstadt.de

Abstract

Object class detection in scenes of realistic complexity remains a challenging task in computer vision. Most recent approaches focus on a single and general model for object class detection. However, in particular in the context of image sequences, it may be advantageous to adapt the general model to a more object–instance specific model in order to detect this particular object reliably within the image sequence. In this work we present a generative object model that is capable to scale from a general object class model to a more specific object–instance model. This allows to detect class instances as well as to distinguish between individual object instances reliably. We experimentally evaluate the performance of the proposed system on both still images and image sequences.

1. Introduction

The ability to detect objects and pedestrians in still images and image sequences is key to a variety of important applications such as surveillance, image and video indexing, intelligent vehicles or robotics. Most research in the area has focused on approaches to effectively model intra–class variation to generalize well across object class instances. Tremendous progress has been made for object as well as pedestrian detection [19, 25, 5, 11, 9, 24, 17, 15, 10, 4, 13, 21].

An open problem, however, is to detect multiple objects and pedestrians in crowded scenes where pedestrians might be significantly occluded over longer periods of time. Traditionally, approaches in this area have been formulated as tracking problems [6, 27, 26, 23, 18] due to the importance of temporal consistency. Quite interestingly many approaches in this area rely on simple object and pedestrian models (i.e. color histograms) suggesting that their effectiveness mostly comes from sophisticated Bayesian and temporal inference mechanisms or from the use of multiple cameras [3, 2, 16, 7, 14]

This paper follows a quite different route by starting from a general pedestrian detection model [10] that is capable to detect and segment pedestrians in images of crowded scenes. In order to handle significant occlusion over longer periods of time we aim to re-detect individual pedestrians previously seen within an image sequence. For this, the general pedestrian model is specialized for the detection of individual pedestrians. As we will discuss below the specialized models leverage e.g. from the segmentation ability of the general pedestrian model to reason about partial occlusions in crowded scenes. These individualized pedestrian models taken together with a simple temporal continuity model allow then to effectively detect multiple pedestrians in crowded scenes despite significant and longer partial occlusions.

The main contribution of the paper therefore is a unified object model that is scalable from a general object–class model to a more specialized and even individualized object–instance model. To learn a robust and accurate model of an individual object or pedestrian from a small number of detections is clearly challenging. Rather than to learn a model for each individual pedestrian from scratch we specialize the general pedestrian model to the individual. The individualized models thereby preserve properties of the more general model such as the segmentation ability and the general pedestrian appearance codebook. To achieve this we first (section 3) extend the original ISM-approach [9] e.g. by incorporating codebook priors. This enables robust learning and the specialization of individualized pedestrian models from few training samples. The proposed extensions are experimentally compared to previous published results on challenging data–sets showing the validity of the approach. Section 4 then applies this approach to image sequences with significant occlusion over longer periods of time. Experimental results show that these specific models can be used to increase both precision and recall of the detection.

2. Original Implicit Shape Model Approach

The Implicit Shape Model (ISM) developed by Leibe and Schiele [9] is a generative model for general object detection. It has been applied to a variety of object cat-

egories including cars, motorbikes, cows and pedestrians. For pedestrians a number of extensions [10, 21, 22] have been proposed, which exploit the nature of this object category by incorporating knowledge about pedestrian articulation. In this paper we propose to improve and extend the probabilistic modeling of the original approach. As will be shown later this allows not only to train general models for pedestrian detection but also to train specialized models for individual pedestrians. Before introducing the extensions in section 3, the following briefly explains the steps involved in learning an object model in the original ISM framework.

Appearance Codebook. A visual vocabulary, referred to as appearance codebook, is used to describe common object features or parts of an object class. To learn the appearance codebook a scale-invariant interest point detector (Hessian-Laplace [12]) is applied to each training image and local image descriptors (Shape-Context [1, 12]) are extracted around them. These image descriptors are subsequently clustered with an agglomerative clustering scheme.

Spatial Occurrence Distributions. Once an appearance codebook is learned for an object class, *separate* spatial occurrence distributions for each codebook entry c_i are learned. During a second run over the training images the codebook is matched to the training examples and occurrence locations (x-, y-position, scale) for the codebook entries are recorded.

Recognition. During recognition, the same feature extraction procedure is applied to obtain a set of local image descriptors e at various scales on the test image. A local image descriptor e extracted at the absolute image coordinates ℓ is compared to the appearance codebook. A descriptor may have multiple matching codebook entries c_i. Let $p(c_i|e)$ denote the matching probability. For each possible codebook match votes are cast for different object centers λ_x, λ_y and scales λ_σ according to the individual occurrence distributions with $\lambda = (\lambda_x, \lambda_y, \lambda_\sigma)$. Each vote has the weight $P(o_n, \lambda|c_i, \ell) \cdot p(c_i|e)$. A descriptor's contribution to the recognition process can therefore be expressed by the following marginalization:

$$P(o_n, \lambda|e, \ell) = \sum_{c_i} P(o_n, \lambda|c_i, e, \ell)p(c_i|e) \quad (1)$$

$$= \sum_{c_i} P(o_n, \lambda|c_i, \ell)p(c_i|e) \quad (2)$$

The overall object probability at position λ is obtained by summing over all extracted descriptors e_k:

$$P(o_n, \lambda) = \sum_{k} P(o_n, \lambda|e_k, \ell_k) \quad (3)$$

Maximum search is accomplished by Mean-Shift Mode Estimation with a scale-adaptive kernel K [8].

Segmentation and MDL-based verification. Next to object localization, a pixel-wise segmentation can be inferred for each hypothesis. Finally, a Minimum Description Length (MDL) based verification step is applied in order to disambiguate overlapping hypotheses. As has been previously shown the segmentation step in combination with the MDL selection mechanism significantly improves detection performance and allows a pixel–level reasoning about occlusions. For the computational details please refer to [9].

3. Extensions to the Implicit Shape Model

The Implicit Shape Model has shown state-of-the-art performance for the detection of pedestrians in images of crowded scenes. In order to further improve its performance we extend the ISM formulation in various ways. The general aim is to derive a unified ISM formulation that on the one hand allows to train a general pedestrian detection model and on the other hand to specialize this general pedestrian model to enable robust detection of individual pedestrians in image sequences. More specifically this section introduces a novel probabilistic modeling scheme as the basis for the unified formulation and discusses the necessary steps to make detection more reliable and to better exploit the information available in the training data. As will be seen in the experiments this allows not only to train individualized pedestrian models but also also improves the results for general pedestrian detections w.r.t. the original ISM formulation.

Appearance Codebook. As in the original approach we learn an appearance codebook by agglomerative clustering. To only keep codebook entries that are representative for an object class such as pedestrians we discard codebook entries, which very rarely match to the training images.

Global Spatial Occurrence Distribution. Instead of learning individual occurrence distributions for each codebook entry separately as in the original ISM, we propose to learn a joint occurrence distribution for the entire appearance codebook. A joint occurrence distribution has several advantages. Firstly, individual occurrence distributions as used by Leibe & Schiele assume, that all codebook entries are equally important. However, there are some codebook entries which are more typical for an object class than others. For cars, for example, wheel features are crucial for reliable detection.

Secondly, even if a codebook entry occurs frequently enough on pedestrians in general, when training an individualized pedestrian model from a small number of training samples we may have insufficient statistics resulting in degenerated occurrence distributions. For example, a codebook entry occurring only (even several times) at a single location in the training data, concentrates its entire probability mass on a single point. During recognition this can have the effect, that an hypothesis is dominated by such a degenerated distribution, yielding false positive detections with very high scores. This effect is particularly likely when

training from as few as 2 or 3 training images as done below.

Finally, a joint occurrence distribution ensures that the model is normalized as a whole instead of separately for each codebook entry. As a result the recognition scores become more comparable across (individual or separate) object models.

The following explains in more detail how the global occurrence distributions $P_{occ}(o_n, \lambda, c_i)$ are learned on the training set. $P_{occ}(o_n, \lambda, c_i)$ denotes the probability, that codebook entry c_i is observed and that the object center is at position λ relative to codebook entry c_i. For the derivation we assume that a general pedestrian appearance codebook $C = (c_1, \ldots, c_n)$ has been learned on a set of pedestrian images. The occurrence distributions themselves can be learned on the same set of pedestrian images (to obtain a general pedestrian model) or on an independent set of images for example from an individual pedestrian.

Let e be an image descriptor extracted at location $-\lambda$ on the training set (the location is recorded with respect to the object center). We compare e to each of the codebook entries, with $p(c_i|e)$ denoting probability that e is associated with entry c_i. The descriptor's contribution to the global occurrence distribution is distributed over the codebook dimension of $P_{occ}(o_n, \lambda, c_i)$ according to the probabilities $p(c_i|e)$. As a result, each training feature has the same influence on the final object model. The model therefore cannot be dominated by rare occurrences or even outliers and learns a better representation of the mean structure of the object class. Additionally, the frequency information of the codebook entries is retained in the model. One can also think of this process as the introduction of priors $p(c_i)$ for each codebook entry ($P_{occ}(o_n, \lambda, c_i) = P(o_n, \lambda|c_i)p(c_i)$). Where the priors are determined by the occurrence frequency. Note, that the codebook priors are learned on the individual training samples whereas the codebook entries themselves may be learned on a larger set of pedestrian images.

Recognition. After having learned the new object model $M = (A, P_{occ})$ consisting of an appearance codebook A and a global occurrence distribution P_{occ}, we apply the same scale-invariant interest point detector on a test image to obtain local image descriptors at various scales.

Again, let e denote a local image descriptor extracted at the absolute image coordinates ℓ. Image descriptor e is matched to each entry of our appearance codebook. For each matching codebook entry or object part we cast votes for different object centers λ in a continuous 3D voting space according to the recorded *global* occurrence distribution P_{occ}. We refer to the set of codebook entries matching to image descriptor e as $M(e)$.

The contribution of a descriptor e can then be expressed by the following equation:

$$P(o_n, \lambda|e, \ell) = \sum_{c_i} P_{occ}(o_n, \lambda, c_i|e, \ell) \quad (4)$$

$$= \sum_{c_i \in M(e)} P_{occ}(o_n, \lambda, c_i|\ell) \quad (5)$$

$$= \sum_{c_i \in M(e)} P_{occ}(o_n, \lambda - \ell, c_i) \quad (6)$$

Note, that a vote $Pocc(o_n, \lambda - \ell, c_i)$ represents evidence for a certain codebook entry c_i to be present at location ℓ in the test image. The are two main differences to the original ISM recognition procedure. As pointed out before we use the joint occurrence distribution P_{occ} rather then the individual codebook distributions. The other difference is that previously each feature's contribution was distributed across multiple codebook entries based on the matching probability $p(c_i|e)$. Whereas here the features activate all matching codebook entries completely. As we will see below these differences will result in a better detection performance of the general pedestrian detection model and will also allow to train individualized pedestrian models.

The object probability for a location λ in the test image can then be computed by:

$$P(o_n, \lambda) = \sum_k P(o_n, \lambda|e_k, \ell_k) \quad (7)$$

$$= \sum_k \sum_{c_i \in M(e_k)} P_{occ}(o_n, \lambda - \ell_k, c_i) \quad (8)$$

Thus, an object hypothesis is the summation of probabilistic votes pointing to the same object center. Since each vote represents evidence for a codebook entry, a hypothesis can be considered to be a collection of codebook entries appearing at certain positions in the test image.

For the maximum search we use Mean-Shift Mode Estimation with a scale-adapted kernel volume.

$$\hat{p}(o_n, \lambda) = \frac{1}{nh(\lambda)^d} \sum_k \sum_j p(o_n, \lambda_j|e_k, \ell_k) K(\frac{\lambda - \lambda_j}{h(\lambda)}) \quad (9)$$

The kernel volume h is chosen in a way, that detection is robust to center point variations in the training and test data. However, integrating over the kernel volume has the effect that parts of the object model are explained multiple times. This happens, when similar local appearances are found very close to one another (with respect to x-,y-coordinates and scale) in the test image. Consider, for example, a codebook entry c_i occurring at position $-\lambda$ in the training set. If two descriptors e_1 and e_2 with similar appearance are found close to one another at ℓ_1 and ℓ_2 in the test image, their contributions $P_{occ}(o_n, \lambda - \ell_1, c_i)$ and $P_{occ}(o_n, \lambda - \ell_2, c_i)$ will be in the same kernel volume. To

avoid that the contribution is accounted for multiple times we remove the redundant evidence from the kernel volume during the maximum search.

The remaining votes are then back-projected to the image and a pixel-wise segmentation mask is inferred for the object hypothesis. Note that this can be done only because we use a general pedestrian codebook for which we have learned the respective segmentations as well. When we have a few detections for an individual pedestrian in an image sequence and we want to learn a specialized model for this individual we cannot assume highly accurate segmentations for those few training samples. Therefore we leverage from the segmentation ability of the general pedestrian appearance codebook to be used for the individualized models. As will be seen later this leads to segmentations for individuals that can be used again for pixel-wise occlusion reasoning. Finally we apply the MDL based verification stage to disambiguate overlapping object detections. Note also, that the final MDL verification step helps to decorrelate the influences of overlapping descriptors.

3.1. Evaluation of the General Object Model

In order to evaluate the proposed extensions to the object model, we applied it to three challenging pedestrian data sets. These data sets range from single-person side-view images to multi-person and multi-viewpoint detection in the presence of clutter and occlusion.

On test set A we evaluate the detection performance, when people are fully visible. This test set contains 181 side-views of pedestrians in different articulations and with a wide range of different clothing (see images in the lower left corner of Figure 1).

Figure 1 (upper left corner) depicts the obtained result. We compare the new approach both to the results of the original authors, as well as to the Histogram of Oriented Gradients (HOG) detector of Dalal & Triggs [4], which is based on a global descriptor instead of local image features.

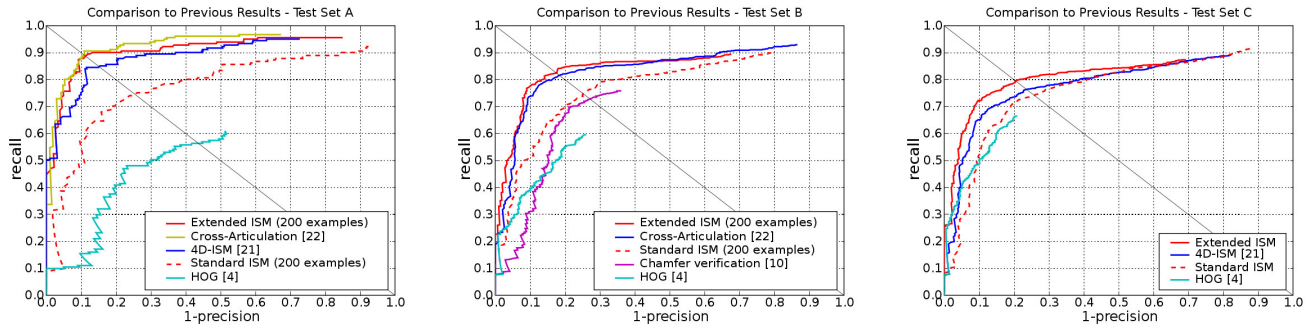
As can be seen, on this data set the HOG detector performs rather poorly with an equal error rate (EER) of only 57%. This is, on the one hand, due to the fact, that the data set is quite challenging. On the other hand, a pretrained binary of the HOG detector was used, which was optimized for multi-viewpoint detection. The original ISM approach achieves an EER of 74%. Our new model (red curve), which is based on a global occurrence distribution outperforms these results by 14% and reaches an EER of 88%. This is a significant improvement and shows the potential of the novel object model.

Note, that various extensions have been proposed that can further improve the performance of the original ISM approach. The 4D-ISM [21] yields an EER of 85% and *Cross-Articulation Learning* [22] 89% respectively. However, these extensions exploit explicit knowledge about pedestrian articulations. When using the *Cross-Articulation Learning* approach on top of the newly proposed object model, we can further improve the results. However, the improvement is less significant and performance improves only slightly in terms of precision compared to [22]. As we want to use the proposed extensions to learn from as few as 2 or 3 training samples it is not clear how these further extensions could be incorporated to train individualized detection models.

To confirm the suitability of the new object model in the presence of significant occlusions and overlapping people, we use the *Crowded Scene* data set from [10]. We refer to the data set in this paper as test set B. It contains 206 images with a total number of 595 annotated pedestrians. On this data the HOG detector attains a recall of 60% for a precision value of 75%. Note, that a higher recall could not be achieved with this detector, since the provided binary has a fixed confidence threshold. The original ISM approach yields an EER of 73%. The newly proposed method increases the detection performance to 82%. Again, this is a significant improvement. On this data set we even outperform the the more elaborate *Cross-Articulation Learning* approach from [22], which achieves only an EER of 81%.

Finally, we tested the new approach on the multi-viewpoint test set C [21]. This test set includes not only overlapping and occluded pedestrians, it also shows them from different viewpoints. The total number of images is 279 with 847 annotated pedestrians. The HOG detector and the original ISM approach have similar detection performance, with the HOG detector performing better in the first part of the curve and slightly worse in the second part. The original ISM's EER is at 74%. The new approach performs significantly better along the whole precision-recall curve, achieving an EER of close to 80%. It also outperforms the 4D-ISM approach [21].

As our experiments have shown, the new object model yields significant performance improvements compared to the original ISM model on a variety of databases. The original ISM-approach has been improved by explicitly incorporating articulation information. These improvements resulted in the best results so far reported on these databases. Interestingly, the novel approach proposed in this paper achieves comparable or even superior detection rates without the explicit use of articulation information. Considering the difficulty of the databases this shows that the new model makes pedestrian detection very robust even in the presence of clutter and occlusion. In order to stress this, Figure 1 (second row) depicts some example detections of our system. The following section now discusses that the novel approach lends itself to robustly learn specialized pedestrian model from as few as 2 or 3 training samples.



(a) Test Set A (b) Test Set B (c) Test Set C

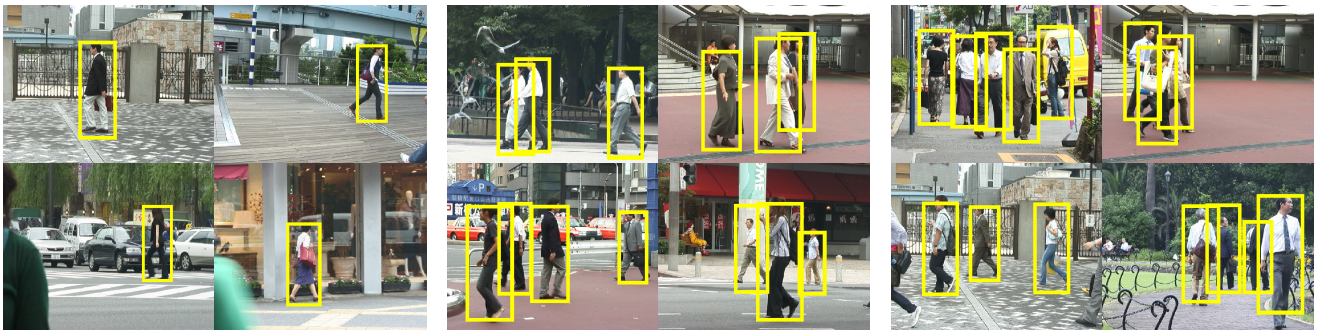


Figure 1. *First row: Recall precision curves, which compare our detection performance to results from other approaches for the different test sets. Second row: Example detections for test set A, B and the multi-viewpoint data set C.*

4. Specialized Object Models

The proposed probabilistic formulation of the object model enables new possibilities and applications. In this section, we will explain, how a general object model can be specialized to a single pedestrian instance. Hereby, we are able to leverage from both the general pedestrian appearance codebook and the segmentation abilities of the general model.

In the first step, however, we would like to experimentally verify, that, given the new formulation, learning from a small number of training examples is feasible.

4.1. Varying Training Set Size

In this experiment we gradually decrease the number of training examples. From the 200 original images in the training set, we first randomly select a set of 50 and finally a set of only 10 pedestrians. We compare the obtained results to those of the original ISM approach.

Figure 2 depicts the corresponding results. As can be seen, the recognition rates drop when the number of training examples is reduced. The EER for 200 training examples is 89%, for 50 training examples 77% and 62% for 10 examples. Of course, this was to be expected. However, even with as little as 10 training examples (blue curve in Figure 2), we achieve reasonable detection results with our new model. The detection recall at the EER always exceeds 60%, which is remarkable. In fact, it is even better than the state-of-the-art HOG detector. Compared to the original ISM approach, the detection precision is improved significantly when learning from only 10 examples. This can be explained by the codebook priors, which successfully reduce the influence of degenerated occurrence distributions. At a recall level of 50%, 85% of our hypotheses are correct, while for the original ISM this is only true for approximately 50%.

4.2. Instance-Specific Models

General object detectors model a complete object category. That is why they have to be able to cope with large intra–class variations. Instance-specific models, on the other hand, can be directly adapted to an individual appearance. Their focus is it to successfully detect the same instance, as well as to distinguish it from different pedestrian instances. Thus, it is possible to re-detect pedestrians in an image sequence despite significant and longer partial occlusions. One can imagine to train a specialized model for an individual when a sufficient number of training samples is available. However, the real challenge is to learn a specialized model from as few training instances as possible.

Let us now consider the details involved in deriving a

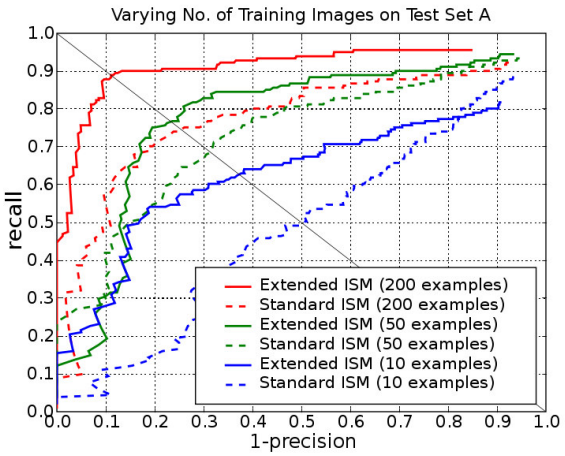


Figure 2. *Recognition performance for varying number of training images on side-view pedestrians (Test Set A).*

specific model for a detected pedestrian, based on detection hypotheses from the generic object model. A detection hypothesis H is obtained by a summation over the contributions of the individual image descriptors (see equation 8). We can rewrite equation 8 in the following manner:

$$P(o_n, \lambda) = \sum_k \sum_{c_i \in M(e_k)} P_{occ}(o_n, \lambda - \ell_k, c_i) \quad (10)$$

$$= \sum_{(c_i, \ell_k) \in H} P_{occ}(o_n, \lambda - \ell_k, c_i) \quad (11)$$

where (c_i, ℓ_k) are pairs with c_i matching to the test image descriptor e_k at position ℓ_k.

The equation expresses that an object hypothesis is a sum of samples from the global occurrence distribution P_{occ}. The samples are drawn from the object parts c_i which have been found in the test image. In other words, the terms $P_{occ}(o_n, \lambda - \ell_k, c_i)$ in equation 8 denote the probability, that codebook entry c_i is observed at position $\lambda - \ell_k$ in the test image.

We can now consider the test hypothesis to be another training example. We know its object center (λ) and we know which codebook entries have occurred at which position on this object instance. This information is enough to build a new object model for exactly this hypothesis. In fact, we can reuse the samples from general object model directly to derive a specialized occurrence distribution P_s

$$P_s(o_n, \ell_k - \lambda, c_i) = \frac{p(c_i)}{Z} \sum_{\ell_k \in H} P_{occ}(o_n, \lambda - \ell_k, c_i)$$

where Z is a normalization factor, which ensures, that the occurrence distribution of the instance-specific model is normalized.

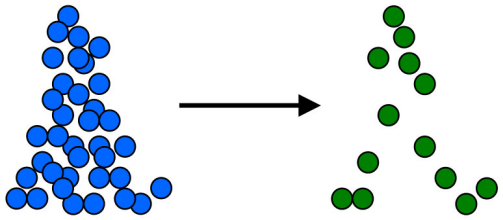


Figure 3. *The general occurrence distribution (left) contains contributions from various people and articulations. (The blue circles denote codebook entry occurrences relative to the object center). The instance-specific model (right) is a subset of the general model.*

Figure 3 illustrates the process. A sub-part of the general occurrence distribution (visualized by the blue circles) is used as the new specific occurrence distribution.

In this manner, instance-specific object models can be created from the general object model on-the-fly. The resulting models benefit from the general model in two ways. Firstly, they are based on the same general appearance codebook. Secondly, they inherit the segmentation abilities of the general model. Intuitively, the general model has for each codebook entry occurring at a certain location an associated segmentation mask (for further details please refer to [9]). These masks can be passed to the specific model, thus allowing to build a top-down segmentation from hypotheses of the specific model. Figure 4 shows example segmentations for a specific pedestrian model, which was initialized when the person was standing upright with closed legs (rightmost image). As can be seen, the person can be successfully detected in subsequent frames. However, as the appearance of the person changes due to articulations changes, the detection relies almost exclusively on features on the upper body when the leg articulations change significantly (middle). When a similar leg articulation as during training appears the detection can again use the respective leg parts (left). The inclusion of articulation information therefore has the potential to improve also the detection of individual pedestrians which is left for future research.

4.3. Instance-Specific Evaluation

In order to show the principal effectiveness of instance-specific object models, we learn 5 different object-specific models for the pedestrian present in an image sequence. The image sequence is recorded in a challenging setting, with pedestrians entering from the left and right sides and crossing each other. Therefore most pedestrians are partially or even completely occluded during the sequence. Figure 5 shows some example images from the image sequence. For ground-truth annotations of the sequence, a pedestrian was considered, when it was approximately 20% visible. As instance-specific models are sensitive to articulation changes, we initialized the models from three general detections in consecutive image frames. This ensures,

Figure 4. *Example segmentations for an instance-specific object model. The model was initialized when the legs were closed, thus detection of the legs fails, when the person is making a large step. However the evidence, from the upper body is sufficient to successfully detect the person.*

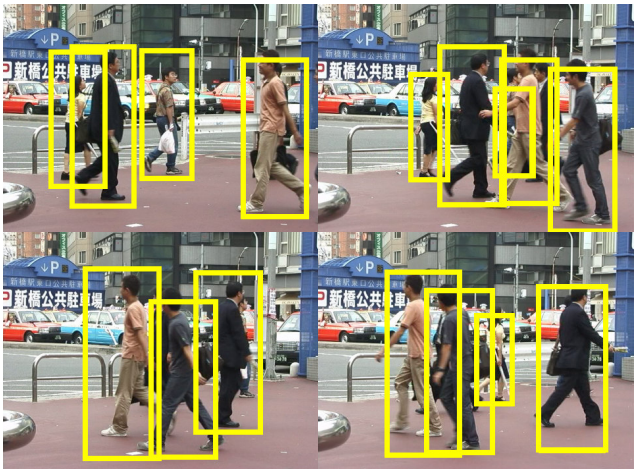


Figure 5. *Example detections on a challenging image sequence, with cluttered background and heavy occlusion. Upper left: 4 of 4 pedestrians detected. Upper right: 4 of 5 pedestrians detected. Lower left: 3 of 4 pedestrians are detected. Lower right: 4 of 5 pedestrians detected.*

that the resulting models are more robust w.r.t different body poses.

As an initial test, we computed how well the specific models are able to distinguish the different persons in the sequence. Therefore, we applied each instance model separately to each frame in the image sequences. Then, for each ground-truth annotation, we determine by which object model the pedestrian was explained best. Table 1 shows the respective results in terms of a confusion matrix.

Person $P2$ is, for example, recognized in 72% of the cases by the correct model, while in 7% of the cases the object model from person $P3$ yielded better detection scores. Considering that we have a 5 class problem and that pedestrians are often partly occluded, these are respectable results. Please also keep in mind, that our models should not only achieve good detection precision, but also a high recall. Our evaluation has shown that, even though the specific models are learned from only three consecutive frames, they are flexible enough to successfully detect the person almost throughout the sequence.

	P1	P2	P3	P4	P5
P1	**68%**	14%	–	7%	11%
P2	14%	**72%**	7%	–	7%
P3	12%	–	**88%**	–	–
P4	12%	4%	4%	**72%**	8%
P5	9%	13%	13%	9%	**55%**

Table 1. Confusion matrix between 5 instance-specific object models on an image sequence with significant partial occlusion.

Finally, we show how the instance-specific models can be used to increase detection robustness. Since we can accurately distinguish individual persons in the image sequence, it is possible to follow a detection hypothesis based on its location even when people are overlapping. When a person is fully occluded, we can recover based on its specific appearance as soon as it becomes visible again. With a general object model alone, this would not be feasible.

When performing object detection with the generic object model on the described pedestrian sequence, an EER of 79% is reached (see Figure 6). The maximum recall is approximately 82%. Note, that this value is rather low as pedestrians are sometimes 80% occluded. When following pedestrian hypotheses based on the 5 instance–specific models, a recall of 71% is obtained. However, the detection precision is significantly improved, with only two false positive detections at over 70% recall (yellow curve). In order to obtain higher recall values the instance–specific models are not general enough.

Fortunately, the specific object models tell us exactly, which path a person has taken. Thus, we can fill the missing detections with detections from the general object model. In this way we combine, instance-specific detections and general detections. As can be seen in Figure 6 (red curve) this can increase detection recall to a value of 86%. Consequently a combined detection system based on generic and specific object models can successfully improve detection results on image sequences. Figure 5 shows some example detections of the final system. A video of the system will be available through the supplementary material.

5. Conclusion

We presented a unified probabilistic formulation for a model that is scalable from general object–class detection to specific object–instance detection. Our experimental evaluation on different pedestrian image databases has shown, that the general object detection performance has been sig-

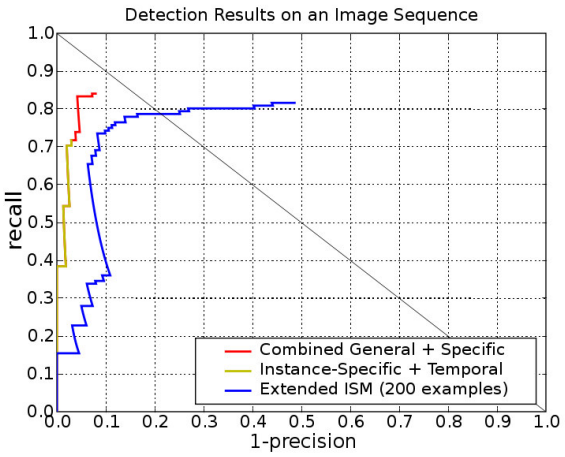


Figure 6. *Recognition performance on an image sequence with a large number of occlusions. Note, that the instance-specific results (yellow) overlay the combined approach (red) in the first part of the curve.*

nificantly increased compared to the original ISM approach and w.r.t to other state-of-the-art detection systems.

We developed a method to learn instance–specific object models from detection hypotheses of the general object model. Thereby, the instance–specific models share the same codebook representation and can leverage from the segmentation ability of the general pedestrian model. The conducted experiments have shown, that the instance–specific models are powerful enough to detect the object of interest in an image and to distinguish between other objects of the same category. When combining both general detections and instance-specific detection, we can successfully increase the robustness of the detection system.

In the future we will address the open issue on how to include articulation changes for the instance–specific models. Updating the model over time for example would make it possible to learn the specific walking cycle of a pedestrian.

Acknowledgements: This work has been funded, in part, by the EU project CoSy (IST-2002-004250) and Toyota Motor Europe.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 2002.

[2] J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In *CVPR*, 2006.

[3] Q. Cai and J. Aggarwal. Automatic tracking of human motion in indoor scenes across multiple synchronized video streams. In *ICCV*, 1998.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] R. Fergus, A. Zisserman, and P. Perona. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.

[6] M. Isard and J. MacCormick. Bramble: A bayesian multiple-blob tracker. In *ICCV*, 2001.

[7] S. Khan, O. Javed, and M. Shah. Tracking in uncalibrated cameras with overlapping field of view. In *PETS*, 2001.

[8] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[9] B. Leibe and B. Schiele. Scale invariant object categorization using a scale-adaptive mean-shift search. In *DAGM*, 2004.

[10] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, 2005.

[11] F.-F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshop On Generative-Model Based Vision*, 2004.

[12] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 2005.

[13] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV*, 2004.

[14] A. Mittal and L. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. In *IJCV*, 2003.

[15] A. Mohan, C. Papageorgiu, and T. Poggio. Example-based object detection in images by componentes. In *PAMI*, 2001.

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

[17] A. Opelt, A. Pinz, and A. Zisserman. A boundary-fragment-model for object detection. In *ECCV*, 2006.

[18] J. R. Peter, H. Tu, and N. Krahnstoever. Simultaneous estimation of segmentation and shape. In *CVPR*, 2005.

[19] H. Schneiderman and T. Kanade. A statistical method of 3d object detection applied to faces and cars. In *CVPR*, 2000.

[20] E. Seemann, B. Leibe, K. Mikolajczyk, and B. Schiele. An evaluation of local shape-based features for pedestrian detection. In *BMVC*, 2005.

[21] E. Seemann, B. Leibe, and B. Schiele. Multi-aspect detection of articulated objects. In *CVPR*, 2006.

[22] E. Seemann and B. Schiele. Cross-articulation learning for robust detection of pedestrians. In *DAGM*, 2006.

[23] K. Smith, D. G.-Perez, and J.-M. Odobez. Using particles to track varying numbers of interacting people. In *CVPR*, 2005.

[24] A. Torralba, K. Murphy, and W. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*, 2004.

[25] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *ICCV*, 2003.

[26] B. Wu and R. Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. In *CVPR*, 2006.

[27] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *CVPR*, 2004.