Multi-Aspect Detection of Articulated Objects

Edgar Seemann, Bastian Leibe and Bernt Schiele
Computer Science Department
Darmstadt University of Technology
{lastname}@mis.tu-darmstadt.de

Abstract

A wide range of methods have been proposed to detect and recognize objects. However, effective and efficient multi-viewpoint detection of objects is still in its infancy, since most current approaches can only handle single viewpoints or aspects. This paper proposes a general approach for multi-aspect detection of objects. As the running example for detection we use pedestrians, which add another difficulty to the problem, namely human body articulations. Global appearance changes caused by different articulations and viewpoints of pedestrians are handled in a unified manner by a generalization of the Implicit Shape Model [5]. An important property of this new approach is to share local appearance across different articulations and viewpoints, therefore requiring relatively few training samples. The effectiveness of the approach is shown and compared to previous approaches on two datasets containing pedestrians with different articulations and from multiple viewpoints.

1. Introduction

Detecting instances of an object category such as pedestrians from single still images has been an active research topic for a number of years [4, 10, 9, 15, 7, 6, 3, 16]. Many of the approaches use appearance based models and good results have been reported on various databases. While the approaches are typically fast, most of them have been only trained and used for single aspects or viewpoints of pedestrians such as side-views or front/back-views with two notable exceptions [15, 3].

The standard approach to multi-viewpoint object detection is to use several detectors running in parallel and combine their outputs via a complex arbitration scheme [11]. The main drawbacks of this approach are the need for a complex arbitration logic and for larger amounts of training data. In addition, it is problematic how a discriminative classifier for multiple (often similar and correlated) viewpoints of the same object can be trained. Interestingly, recent work [14] has shown that the individual detectors' discriminance can be increased and the training data can be more efficiently used when features are shared between detectors.

Even when features are shared between different aspects, current approaches typically require the training aspects to be annotated manually. While this annotation step is still feasible for rigid object categories, it becomes problematic when dealing with articulated categories such as pedestrians. The combination of viewpoint and articulation changes makes it difficult to discretize training views into consistent sets, let alone to decide how this discretization roster should be set up.

Instead of specifying the training aspects manually, it becomes therefore desirable to deal with various global appearance changes caused e.g. by human body articulations, viewpoint changes or object deformations in a consistent manner. Ideally, those consistent appearance changes are learned automatically from training data, e.g. by clustering coherent views. [4] pursues such an approach for detecting pedestrian side views by clustering silhouettes and storing them in a tree to speed up Chamfer matching. However, global approaches, such as the Chamfer matching method, are not robust to partial occlusion and local deformations. Local approaches can in general deal better with partial occlusions and local deformations but are less suitable to guarantee global consistency. Therefore it is typically difficult for local approaches to handle or even reliably estimate multiple viewpoints.

The method presented in this paper builds closely on the approach developed in [6]. In that approach the Implicit Shape Model (ISM) as a local approach is combined with a global verification stage based on silhouettes. The global, silhouette-based verification step essentially allows to enforce global consistency of object hypotheses generated from local evidence. While the principal effectiveness of the approach has been shown, the global nature of the verification step makes it difficult to deal with partial occlusion. Rather than to rely on a purely global verification stage the approach proposed here uses a semi-local (or semi-global) verification stage for different articulations and viewpoints in a unified manner. The promise is that this semi-local verification is more discriminative than a purely local approach while being more robust than a purely global approach. Interestingly,

the verification stage can be scaled to behave more locally or more globally depending on the object category at hand.

The main contributions of the paper are the following. First, the paper introduces a unified approach for multi-viewpoint and multi-articulation detection of pedestrians. Second, through local appearance sharing across articulations and viewpoints we can effectively learn a multi-viewpoint and multi-articulation object model from relatively few training samples. In a sense this algorithm generalizes the idea of sharing of features [14] to the sharing of local appearance across object instances, viewpoints and articulations. Third, the detection algorithm combines the robustness of local approaches to partial occlusion and to local deformations with the advantages of global consistency verification. The semi-local nature of this approach can be scaled either to behave more locally or globally e.g. depending on the amount of global deformations of the respective object class

After a brief review of the original ISM recognition approach (see Section 2) the new extended 4D-ISM approach is introduced in section Section 3. Section 4 describes experimental results on two challenging datasets containing pedestrians from multiple viewpoints and with different articulations.

2. Standard Implicit Shape Model

ISM Training. An ISM [6] is trained by extracting local features from training examples and modelling their spatial occurrence distributions on the object category. For this, a scale-invariant interest point detector is applied to each training image, and local descriptors are calculated on the extracted regions. Subsequently, the local descriptors are clustered to form a visual vocabulary (or *codebook*) of typical object structures. In a second run over the training data, the spatial distribution of each codebook entry is estimated by recording all matching locations on the training objects. In addition to each occurrence location, a local segmentation mask is stored, which is later used to infer top-down segmentations for detection hypotheses.

ISM Recognition. For recognition, the same feature extraction procedure is applied, and extracted features cast votes for object hypotheses in a probabilistic extension of the Hough transform [6]. Let e be a local descriptor computed at location ℓ. Each of the local descriptors is compared to the codebook and may be matched to several codebook entries. One can think of these matches as multiple valid interpretations I_i for the descriptor, each of which holds with the probability $p(I_i|e)$. Each interpretation then casts votes for different object categories o_n, locations λ_x, λ_y and scales λ_σ according to its learned occurrence distribution $P(o_n, \lambda|I_i, \ell)$ with $\lambda = (\lambda_x, \lambda_y, \lambda_\sigma)$. Thus, any single vote has the weight $P(o_n, \lambda|I_i, \ell)p(I_i|e)$ and the descriptor's contribution to the hypothesis can be expressed by the following marginalization:

$$P(o_n, \lambda|e, \ell) = \sum_i P(o_n, \lambda|I_i, \ell)p(I_i|e, \ell) \quad (1)$$

$$= \sum_i P(\lambda|o_n, I_i, \ell)p(o_n|I_i, \ell)p(I_i|e)$$

$$P(o_n, \lambda) = \sum_k P(o_n, \lambda|e_k, \ell_k) \quad (2)$$

The votes are collected in a continuous 3D voting space, and maxima are found using Mean-Shift Mode Estimation with a scale-adaptive uniform kernel K [5]:

$$\hat{p}(o_n, \lambda) = \frac{1}{nh(\lambda)^d} \sum_k \sum_j p(o_n, \lambda_j|e_k, \ell_k) K(\frac{\lambda - \lambda_j}{h(\lambda)}) \quad (3)$$

The above equations assume statistical independence of the local image regions. While this is not generally valid, it is an approxmation, which works well in practice. Note also, that the final MDL verification step (see below) helps to decorrelate the influences of overlapping descriptors.

Segmentation. Beyond object localization, a segmentation mask can be inferred for each hypothesis. This is accomplished by backprojecting the supporting votes to the image and using the stored local segmentations to infer a pixel-wise segmentation of the object as shown in [5].

Chamfer Verification. In [6], an additional Chamfer verification stage is applied to find a shape template that simultaneously maximizes the Chamfer score and the overlap with the hypothesized segmentation. The overlap is expressed by the Bhattacharyya coefficient [2], which measures the affinity between two distributions. Assuming a uniform distribution for the points inside the shape template s, shifted to location q, its overlap is compared with the hypothesized segmentation Seg:

$$O(q) = \sum_x \sqrt{Seg(x)s(x, q)} \quad (4)$$

and a joint score is computed as a linear combination

$$\text{score} = \alpha \cdot \left(1 - \frac{D_{chamfer}}{\beta}\right) + (1 - \alpha) \cdot O(q) \quad (5)$$

MDL Verification. Finally, a Minimum Description Length (MDL) based verification step is applied in order to disambiguate overlapping hypotheses. This procedure selects the subset of hypotheses which best explains the evidence in the image (see [5] for details).

3. A 4D Implicit Shape Model

For the original ISM-model introduced in the previous section good recognition results have been reported on several object categories including cars, motorbikes, cows and

pedestrians [5, 6]. All of these results, however, have been reported for a single viewpoint of a category. It is therefore unclear how the model performs in a multi-aspect scenario.

This section extends the original model to explicitly handle and estimate viewpoints and articulations of an object category in a consistent manner. Rather than to use a 3D-voting space as in equation 2 the new approach uses a 4D-voting space. The additional dimension summarizes global appearance changes caused for example by human-body articulations, viewpoint changes, or global deformations of the object shape. It is important to note that it is inherently difficult to define a metric on global appearance and shape deformations. The additional dimension therefore consists of an unordered set of discrete global object shapes. The particular set of object shapes is obtained by a clustering scheme. The following introduces the learning of these viewpoint and articulation clusters and then describes the novel 4D-Implicit Shape Model.

Learning viewpoint/articulation clusters. Clearly, manual labelling of aspects in the training data is undesirable, since it is both time consuming and difficult if objects get more complex. Moreover, it's unclear in how many viewpoints or articulations the data should be divided.

Since an object's shape is often a good indication of the current viewpoint and articulation, we automatically learn shape clusters from training data. Each of these clusters corresponds to one articulation and viewpoint. During training we extract object silhouettes or shapes from the training images and cluster these with an agglomerative clustering scheme. The similarity measure between two shapes is based on the Chamfer distance, which enforces global consistency.

As an example Figure 4 shows the articulation clusters for side-view pedestrians generated by this method.

The advantage of this approach is on the one hand, that articulations or viewpoints can be identified without labelling effort, on the other hand it's easy to change the number of clusters by selecting the appropriate level in the clustering hierarchy.

4D Recognition Procedure. Now, having our training data labelled with shape cluster information for different articulations and viewpoints, we add them as a $4th$ dimension to the ISM voting space (see Figure 1). In principle it is possible to extend the probabilistic formulation (eqs. 1 and 2) directly to also incorporate multiple shapes s:

$$P(o_n, \lambda, s|e, l) = \sum_i P(o_n, \lambda, s|I_i, l)p(I_i|e, l) \quad (6)$$

$$= \sum_i P(\lambda, s|o_n, I_i, l)p(o_n|I_i, l)p(I_i|e)$$

$$P(o_n, \lambda, s) \sim \sum_k P(o_n, \lambda, s|e_k, l_k) \quad (7)$$

There are several issues with this formulation. First, it is difficult to estimate the probability density $P(\lambda, s|o_n, I_i, l)$ reliably due to the increased dimensionality, in particular from a relatively small set of data. Second, it would be computationally difficult to perform a maximum search in a 4-dimensional space. Third and quite importantly, since the shape dimension s is neither continuous nor ordered it is not even clear how the maximum search should be formulated. The standard ISM approach applies a Mean-Shift search with a scale-adapted kernel which is no longer feasible for this 4-dimensional case.

Therefore we use the following factorization to obtain a tractable solution:

$$P(o_n, \lambda, s|e, l) = \quad (8)$$
$$\sum_i P(s|\lambda, o_n, I_i, l)P(\lambda|o_n, I_i, l)p(o_n|I_i, l)p(I_i|e)$$

Please note that all but the first term ($P(s|\lambda, o_n, I_i, l)$) are the same as in equation 1. Therefore we can use the following simple yet effective strategy to find the maxima of equation 7. By first searching the K 3D-maxima in equation 2 we can not only reduce the computational complexity but also constrain our search to those areas of the probability density with enough evidence and training data. Choosing K sufficiently large, we can find all maxima with high probability. In practice, the results are rather insensitive to the particular choice of K. For those K maxima we then retrieve the contributing votes and use the following calculation (for simplicity of notation we use $P(s|H) = P(s|\lambda, o_n, I_i, l)$):

$$P(s|H) = \sum_j P(s|c_j, H)p(c_j|H) \quad (9)$$

$$= \sum_j P(s|c_j)p(c_j|H) \quad (10)$$

In this equation c_j corresponds to the individual shapes present in the training data and s is a shape cluster. $P(s|c_j)$ represents the probability that shape c_j is assigned to cluster s. Herefore we can either use hard or soft assignments to shape clusters. If we use hard assignments, we set $P(s|c_j)$ to 1 if c_j is contained in cluster s. For the soft assignments, we use in our experiments the normalized average similarity of a shape to the shape clusters.

By following the above procedure, we can obtain the 4D-maxima of $P(o_n, \lambda, s)$. This means in particular, that the votes corresponding to these maxima conform to a common shape cluster. As a result, the voting scheme produces hypotheses, which have a consistent articulation or viewpoint. An example of this improved consistency is depicted in Figure 2. The two overlapping pedestrians create an hypothesis with a third foot (second image). By considering the shape dimension in the voting procedure, the new approach is able to eliminate this local evidence inconsistent with the estimated shape cluster.

By altering the thresholds of the clustering step, we can adjust the above voting scheme to be more sensitive to local

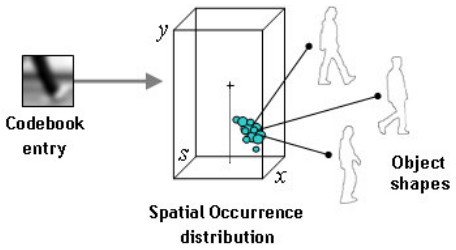

Figure 1. *For each codebook entry, we store the spatial occurrence distribution, as well as the associated shape resulting in a 4-dimensional occurrence space.*

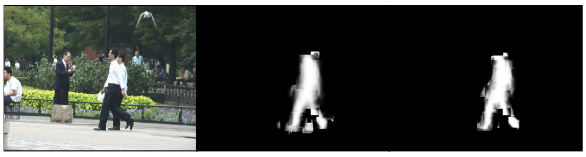

Figure 2. *Overlapping pedestrians can lead to hypothesis with too many extremities. By considering the shape dimension in the voting framework, these influences can be diminished or even eliminated.*

Figure 3. *Example images from training set A*

or global influences. In the extreme case of only one shape cluster, we are back at the original *Standard ISM*, which relies only on local evidence.

4. Experimental Evaluation

The aim of this section is to evaluate the performance of the new 4D recognition scheme proposed in section 3. The main emphasize is on the overall recognition performance by finding hypotheses, which are consistent with the shape clusters. Additionally, we are evaluating the articulation and viewpoint estimation as well.

Two sets of experiments are described. Training and test set A contain pedestrians from a single viewpoint, namely side-views, but different articulations. A particular challenge in this experiment is that the backgrounds and the visual appearance differ considerable between training and test set. See column (a) of Figure 6 for example images of the test set and Figure 3 for examples of the training images.

Training set A contains 210 pedestrians, which are mirrored in order to have the same amount of pedestrians heading left and right. Test set A consists of images from traffic scenes with a total of 181 pedestrians. Training and test set B contain pedestrians from multiple viewpoints. The training set consists of 412 examples. The test set has 279 images with a total of 847 annotated pedestrians (see Figure 9).

For both training sets, segmentation masks are available. These are typically computed from the recorded image sequences with a Grimson-Stauffer background model [13].

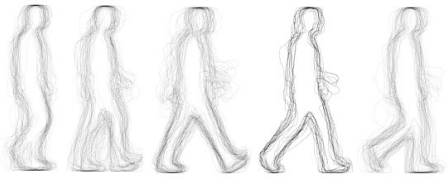

Figure 4. *Automatically found articulation clusters on training set A for the right-left walking direction. There exist an equal number of mirrored clusters.*

From these segmentation mask, we additionally compute the shape silhouettes, which are used for the shape clustering step during model training.

Pedestrians in the test sets are annotated with bounding boxes. Additionally we annotated the articulation and viewpoint. For the viewpoints we use 3 orientations: front-back, side and diagonal. The articulations are annotated according to the clusters which resulted from an agglomerative clustering of the shape silhouettes. This clustering contains 5 articulations from a typical walking cycle and their respective mirrored articulation, which results in a total of 10 clusters (see Figure 4)

4.1. Test Set A - Articulations

Our evaluation on test set A is conducted with regard to pedestrian articulations. For model training we use a modified *Shape Context* descriptor [1, 8], which has been proposed for pedestrian detection in [12]. For comparison we also provide the results obtained with plain image patches as local descriptors.

Figure 5 depicts the respective results. The standard ISM approach using image patches and the Difference-of-Gaussians detector achieves only an equal error rate (EER) of 50%. This is probably due to the large appearance differences between training and test set and the difficult data, which includes heavily cluttered backgrounds. As shown in [6] this performance can be improved by a Chamfer verification stage. For test set A, however, the improvement is moderate.

Using *Shape Context* descriptors along with the Hessian-Laplace detector greatly improves performance to an EER of 74%. *Shape Context* descriptors seem to generalize better, since they operate on edge information only (see also [12]). Further improving these results with a Chamfer verification fails for a number of reasons. Firstly, the remaining object instances are quite challenging as these often correspond to instances with low contrast against the background or significant partial occlusion. Chamfer verification is a global constraint having difficulties with such cases. Particularly if edge structures are not visible at some part of the object. Secondly, *Shape Context* and Chamfer matching both focus on edge information. Thus, there is no additional complementary information which is exploited.

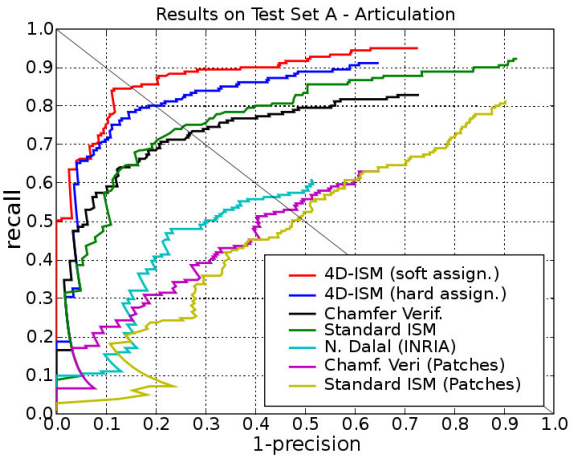

Figure 5. *Recognition performance on side-view pedestrians (Test Set A).*

The proposed 4D-ISM is a semi-local approach and does not suffer from the same drawbacks of the global Chamfer verifi cation. Additionally it exploits consistency information on the articulation clusters not on a single silhouette alone. This information has not been available in the standard ISM approach and results in a signifi cant performance increase. The 4D-ISM achieves up to 85% EER on test set A, which means an improvement of 11% (see Figure 5). Therefore the novel 4D-recognition scheme proves to increase performance both with respect to the original ISM approach as well as the global Chamfer verifi cation scheme.

Figure 6 compares the detection results of Chamfer verifi - cation and the novel 4D-ISM visually by illustrating some of the typical failure cases of the global Chamfer criterion. Row 1 and 2 of the fi gure show how the Chamfer Verifi cation can get distracted by neighboring edge features and shift the hypotheses away from their proper position. Row 4, 5 and 7 depict cases where no training silhouette matches to the test pedestrian. This can be due to different clothing (row 4) or partial occlusion (row $5, 7$). While Chamfer verifi cation fails in these cases, the semi-global 4D-ISM still manages to recover the articulation from local information. In general, the 4D-ISM seems to be much less sensitive to background clutter (see e.g. false positive detections on background in row 4 and 6). This can be explained by the fact, that influences from the background are seldomly consistent with the same shape cluster and can thus be eliminated in the 4D-ISM.

This results in particular in a better detection precision, with the fi rst false positive detections appearing at50% recall (see Figure 5). But also the recognition recall is signifi cantly improved with the new approach.

The results in Figure 5 also show that using *soft assignments* in the 4D voting procedure is superior to *hard assignments*. This was to be expected, since errors in the cluster assignment have a direct impact on the recognition performance. *Soft assignments* avoid a hard decision and feed additional similarity information to the later recognition stages. In this way articulations and viewpoints, which are between the learned clusters are better handled.

(a) (b) (c)

Figure 6. *Articulation estimates containing typical failure cases for Chamfer verification (silhouettes and shape clusters drawn in yellow) – (a) test image, (b) Chamfer verification hypothesis, (c) 4D-ISM articulation estimate*

Figure 7. *Typical failure cases of the 4D-ISM approach. False positives drawn in red.*

Nevertheless there are situations, where even voting with *soft assignments* fails. Figure 7 shows some typical failure cases of the our approach. Mainly, we distinguish two classes of failures. The fi rst class are missed detections due to clothing with poor contrast to the background (see fi rst image of Figure 7). Distracting background structures are responsible for false positive detections, the second class of failures. Even though this effect has been successfully reduced by the proposed approach, false detections occurr if local background structures are particularly strong and numerous over a larger image region.

Finally, we compared the obtained results to the state-of-the-art detector from Dalal&Triggs [3] using the detector available form the author's webpage. We had to adapt the size of the detection bounding boxes, which tend to be quite large and cause our strict evaluation criterion to reject cor-

	Articulation Estimates (column 1-10)									
Annotations (row 1-10)	**63%**	–	5%	–	5%	11%	–	–	16%	–
	8%	**58%**	–	–	8%	–	25%	–	–	–
	–	–	**82%**	–	–	5%	9%	5%	–	–
	–	–	7%	**21%**	36%	–	14%	21%	–	–
	21%	–	–	–	**68%**	–	5%	5%	–	–
	–	–	7%	–	–	**86%**	7%	–	–	–
	6%	–	6%	–	6%	25%	**44%**	6%	–	6%
	–	–	10%	–	–	10%	10%	**70%**	–	–
	–	–	17%	–	–	–	–	17%	**33%**	33%
	–	–	–	–	–	–	17%	8%	–	**75%**

Table 1. Confusion matrix for the articulation estimate (left/upper part for articulations heading right, right/lower part for the mirrored articulations). Rows correspond to the annotations, columns to the estimates.

rect detections as false positives. The detector achieves an EER performance of 57% on test set A. To be fair, it should be mentioned, that our detector was trained in this case on side-views only, whereas their system was built for multi-viewpoint detection. On test set B the comparison will be more meaningful as we also train our detector for multiple viewpoints.

Using the novel 4D recognition scheme we obtain not only a position estimate of the pedestrians, but also an estimate of the human body articulation. In our evaluation, we achieve an articulation recognition rate of 63% for detections at the EER, which is respectable considering that we have a 10 class problem and very difficult background structures. In order to have a proper comparison for these figures, we evaluated the articulation estimate on the hypotheses of the Chamfer verification as well. There an estimation performance of 55% is achieved for the same number of clusters.

Table 1 shows the confusion matrix for the articulation estimates with the 4D-ISM. The upper left part of the matrix corresponds to articulations heading right in the order depicted in Figure 4. The lower right part to their respective mirrored versions. The rows of the matrix correspond to the annotations and the columns to the articulation estimates. Though the estimation is reliable for most of the articulations, it performs poor on some of them. Often, however, the estimate confuses the real articulation with one of the neighboring articulations in a walking cycle (see e.g. row 9, where 17% and 33% of the estimates correspond to the previous and next cluster). Other failure cases include misinterpretation of the walking direction, which can be hard for articulations, that are quite symmetrical (see e.g. row 2, where 25% of the estimates are the correct cluster in the walking cycle, but heading in the opposite direction).

Finally, we observed that the scale estimation for the hypotheses generated by the 4D-ISM is significantly enhanced, too. This indicates that the quality of the hypotheses is improved by the approach. At the EER, the standard ISM has a scale estimation error of 6.1%. The 4D voting scheme results in a scale estimation error of only 4.2%. The reason for this improvement is that the resulting hypotheses of the 4D-ISM are more consistent and less influenced by background structures.

4.2. Test Set B - Viewpoints

On test set B we analyse the performance of the proposed approach for multi-viewpoint images. Again, we apply the standard ISM approach, as well as the new 4D-ISM.

The standard ISM approach achieves already an equal error rate of 74%. Even though all local appearances are incorporated into a common model, which discards any information about viewpoints, the model seems to be able to successfully detect pedestrians in any orientation.

When using 4D-ISM we can further increase this performance 77% EER (see Figure 8). Of special interest is also the fact that the system generates valid hypotheses even if pedestrians are considerably overlapping or occluded. This is one of the strength of this semi-local approach.

Particularly the recognition precision in the first part of the curve benefits from the additional shape dimension with a 16% higher recall for 90% precision. At higher recall values in the second part of the curve, no significant difference can be observed. An improvement in recognition recall would have probably needed a finer grained viewpoint clustering.

In order to compare these results to the state-of-the-art, we applied again the detector of [3]. The performance of this detector on test set B is better than ISM and 4D-ISM for recall values below 40%. Above 40% the proposed 4D-ISM outperforms the detector with, e.g., 8% higher recall for 80% precision. This indicates that our system can generate consistent detection hypothesis for difficult test examples at higher recall values. Especially interesting is, that this performance is achieved by training our detector on only 412 training images. The detector of [3] was trained on 2416 positive examples and 12180 negative examples, which is more than an order of magnitude more. This demonstrates the efficiency of our approach to share features between the different viewpoints and articulations.

As for the viewpoint estimation, an overall accuracy of 71% is achieved for the 3 classes side-view, diagonal and front-back. As can be seen from the confusion matrix in Table 2, there is a bias towards the front-back viewpoint. 23% of the side-view pedestrians and 31% of the pedestrians walking diagonal are recognized as frontal or backwards orientated (last column of the matrix). This seems odd at first, however it can be explained by the fact, that pedestrians with closed legs have very similar shapes in any viewpoint and are hardly distinguishable by the system.

5. Conclusion

In this paper we have introduced a unified approach for multi-aspect object recognition. The approach is able to automatically identify the different aspects in the training data, by clustering the object shapes with an agglomerative clustering scheme. This clustering is used to augment the ISM's

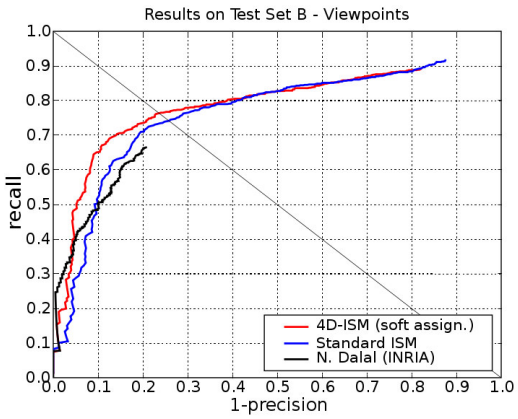

Figure 8. *Recognition performance on multi-view test data (Test Set B). As the provided detector of [3] uses a fixed threshold, we could not determine the later parts of the precision-recall curve of the detector.*

		Viewpoint Est.		
Anno.				
	66%	11%	23%	
	6%	**63%**	31%	
	2%	2%	**95%**	

Table 2. Confusion matrix for viewpoint estimation (side, diagonal, front-back). Rows correspond to the annotations, columns to the estimates.

probabilistic voting scheme with a $4th$ shape dimension. Thus, we can search for object hypotheses in a 4D voting space, which are consistent with the learned aspects. In order to make the search efficient, we proposed to first search in the marginalized 3D space and use these results to infer the maxima of the 4D space.

We have shown the performance of the approach on two challenging data sets with focus to object articulations and multi-viewpoint detection. Our approach outperforms both the original ISM approach of [5, 6], as well as the state-of-the-art detector of [3] on both test sets. Its semi-local nature combines the robustness of local approaches to partial occlusion and local deformations with the advantages of global consistency verification.

By sharing features between object aspects, we can efficiently learn a statistical model from relatively few training examples. In our experiments, we achieve already good recognition rates with $50 - 100$ training examples.

In addition to the improved recognition performance of the approach, the new approach enables the estimation of articulations or viewpoints of a test object. This is, e.g., important in a traffic scenario to predict the direction a pedestrian is heading.

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

Figure 9. *Example detections (in yellow) and false positives (in red) on testset B at the EER.*

References

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

[2] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math Soc.*, 1943.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR'05*, 2005.

[4] D. Gavrila. Multi-feature hierarchical template matching using distance transforms. In *ICPR'98*, volume 1, pages 439–444, 1998.

[5] B. Leibe and B. Schiele. Scale invariant object categorization using a scale-adaptive mean-shift search. In *DAGM'04*, Springer LNCS, Vol. 3175, pages 145–153, Tuebingen, Germany, Aug. 2004.

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

[7] C. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV'04*, LNCS 3021, pages 69–82. Springer, 2004.

[8] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[9] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *TPAMI*, 23(4):349–361, 2001.

[10] C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV'00*, 38(1):15–33, 2000.

[11] H. Schneiderman and T. Kanade. A statistical method of 3d object detection applied to faces and cars. In *CVPR'00*, pages 746–751, 2000.

[12] E. Seemann, B. Leibe, K. Mikolajczyk, and B. Schiele. An evaluation of local shape-based features for pedestrian detection. In *BMVC'05*, 2005.

[13] C. Stauffer and W. Grimson. Adaptive background mixture models for realtime tracking. In *CVPR '99*.

[14] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. 2005.

[15] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *ICCV'03*, pages 734–741, 2003.

[16] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV'05*, 2005.