An Evaluation of Local Shape-Based Features for Pedestrian Detection

Edgar Seemann, Bastian Leibe, Krystian Mikolajczyk, Bernt Schiele
Multimodal Interactive Systems, TU Darmstadt, Germany
{seemann, leibe, kma, schiele}@informatik.tu-darmstadt.de

Abstract

Pedestrian detection in real world scenes is a challenging problem. In recent years a variety of approaches have been proposed, and impressive results have been reported on a variety of databases. This paper systematically evaluates (1) various local shape descriptors, namely Shape Context and Local Chamfer descriptor and (2) four different interest point detectors for the detection of pedestrians. Those results are compared to the standard global Chamfer matching approach. A main result of the paper is that Shape Context trained on real edge images rather than on clean pedestrian silhouettes combined with the Hessian-Laplace detector outperforms all other tested approaches.

1 Introduction

Detecting pedestrians or people has been an active area of research in recent years. Gavrila [5] for example proposes a hierarchy of global pedestrian silhouettes using Chamfer matching and the distance transform to compare the silhouettes with the image content. Papageorgiu et al. [18, 17] use Haar wavelet coefficients to build a global pedestrian model. Zhao and Thorbe [25] perform detection with a neural network and exploit stereo information to pre-segment the image. Viola et al. [24] use simple local features and a boosting scheme to train a cascade of classifiers. They use consecutive frames in an image sequence to detect movement features. Mikolajczyk et al. [13] use several part detectors that are based on local gradient and Laplacian features. Probabilistic co-occurrences of these features help to disambiguate detections. Mori et al. [15] model human body configurations where body part templates are represented by local Shape Context. In later work [16], they apply a normalized cuts segmentation and use shape, shading, and focus cues for retrieving the body parts. Thayananthan et al. [23] compare Shape Context and Chamfer matching for recognizing and localizing gestural hand shapes in cluttered scenes. Forsyth and Fleck [4] introduced the general methodology of body plans for finding people in images. Felzenszwalb and Huttenlocher [3] learn simplistic detectors for individual body parts. Dynamic programming is applied to connect the detected parts to a hierarchy. Ronfard et al. [19] extended this work by using stronger classifiers such as SVMs and RVMs. Mohan and Papageorgiu [14] apply the wavelet-based detectors from [18] to detect body parts and then use body geometry to infer a person's position and pose. Shashua et al. [21] divide a search window into 9 overlapping subregions and classify the regions with a discriminant function. A second-stage classifier based on AdaBoost integrates the obtained results.

It is interesting to note that the variability of approaches ranges from global pedestrian models over part models to models which mostly rely on local features. At the same time the type of features employed ranges from purely silhouette-based to appearance based. Since those approaches are seldomly compared, it is currently quite unclear if global or local models on the one hand and if silhouette or appearance based features on the other hand are more appropriate for pedestrian detection. Traditionally, silhouette-based approaches have been pursued since they appeared more intuitive and appropriate. However the success of more appearance-based approaches questions this intuition.

This paper makes a first step to understand the strengths and weaknesses of various approaches. The main purpose of the paper therefore is a systematic comparison of some novel techniques with existing techniques. The first comparison is between global Chamfer matching to an extension thereof, namely local Chamfer matching. Quite surprisingly, local Chamfer matching clearly outperforms the original global Chamfer matching technique. In a second comparison, local Chamfer matching is compared to local Shape Context, where again local Chamfer matching is found to be the better technique when using silhouette information alone. However when using real edge images for training, local Shape Context outperforms all other tested approaches. Besides this, we also compare four different interest point detectors in combination with the various local features, which results in a clear ranking of the different detectors. All results are validated and replicated on two different training and test sets, suggesting that the obtained results are indeed generalizable.

The paper is structured as follows: Section 2 shortly describes global Chamfer matching as well as Shape Context and local Chamfer descriptors, which are used within the *Implicit Shape Model* framework. Additionally, this section contains descriptions of the properties of the tested interest point detectors. Section 3 explains our test setup for the evaluation and presents the obtained results. A final discussion concludes the paper.

2 System Overview

The main focus of this paper is the comparison of different shape-based detection algorithms for pedestrian detection. A popular detection approach based on global feature is global Chamfer matching [5]. Approaches based on local features can be integrated using a voting framework, which accumulates local evidence to joint hypotheses. The voting framework used in this paper is based on an implementation of the Implicit Shape Model (ISM) [8, 7].

2.1 Global Approach - Global Chamfer Matching

The global Chamfer approach [5] matches object shape silhouettes to image structures. For that purpose, a silhouette is shifted over the image, and a distance $D_{chamfer}(T,l)$ between a silhouette T and the edge image at each image location l is calculated. The distance measure is based on a distance transform DT, which computes for each image pixel the distance to the nearest feature pixel (e.g. edge pixels): $D_{chamfer}(T,l) = \frac{1}{|T|}\sum_{t\in T} DT(t+l)$. The advantage of matching a silhouette with the distance transform rather than the original feature image is that the resulting similarity measure will be smoother [5], which allows to speed up the matching process by employing a hierarchical

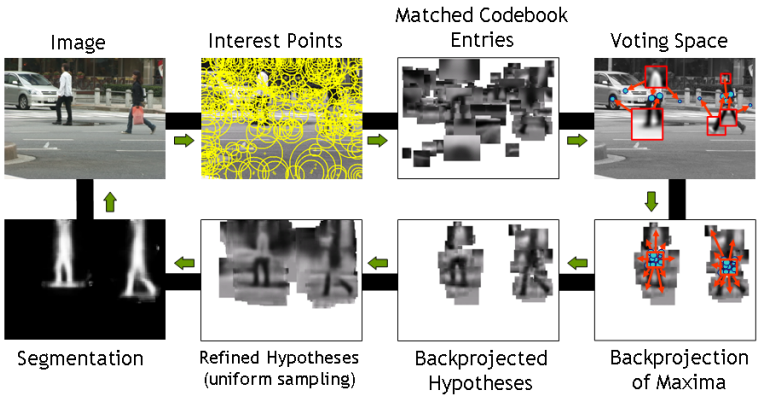


Figure 1: *Recognition procedure for the ISM.*

coarse-to-fine search.

2.2 Local Approach - Implicit Shape Model (ISM)

Model Training. An ISM [8] is trained by extracting local features from the training images and modelling their spatial occurrence distributions on the object. For each training image, an interest point detector D (see section 2.3) is applied, and local features F (see section 2.4) are calculated around the extracted points. Subsequently these local features are clustered to form a visual vocabulary of typical local features, which we call codebook. In a second step, the spatial occurrence distributions on the training data are recorded for each of those typical features. Together with each feature occurrence, a local segmentation mask is stored, which is later used to obtain a top-down segmentation of detection hypotheses.

Hypotheses Generation. During detection, the same feature extraction procedure as in the training step is applied. Codebook entries which match to one of the extracted features cast votes for possible object locations according to the learned occurrence distributions. These votes are collected in a probabilistic Hough voting procedure (see upper part of the loop in Figure 1).

Segmentation and Verification. Beyond the object localization, we can infer a segmentation mask for each hypothesis. This is accomplished by projecting the supporting features of a hypothesis back to the image and using the stored segmentation masks for the local features (see Fig. 3). Finally, a Minimum Description Length (MDL) based verification step is applied in order to disambiguate overlapping hypotheses. For the computational details please refer to [7, 8].

2.3 Interest Point Detectors

Within the ISM approach, an interest point detector D is used to extract feature points. In this paper we use and evaluate four different interest point detectors: Harris, Difference-of-Gaussian, Harris-Laplace, and Hessian-Laplace. The different types of structures extracted by these detectors are visualized in Figure 2.

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

Figure 2: *Interest points (in yellow) on an example image from test set B: (b) Harris (c) DoG (d) Harris-Laplace (e) Hessian-Laplace. Circle diameters are proportional to the scale of the corresponding interest point.*

Harris Detector (Har). The Harris detector [6, 20] was designed to find corner points which are repeatable under translation, image-plane rotation, and noise. The basic idea of this detector is to find image locations where the signal changes in two directions. The most significant signal changes are given by the eigenvectors of the auto-correlation matrix, which takes into account the first derivatives of the signal on a window. Hence, the Harris detector does not respond to straight lines, since the signal changes in a single direction. The Harris detector is the only interest point detector examined in this paper which is not scale-invariant.

Difference-of-Gaussian Detector (DoG). The DoG detector [10] detects stable keypoints across image scales. This is accomplished by searching for scale-space extrema of the difference-of-Gaussian function convolved with the image: $D(x,y,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y)$, where G is a Gaussian function with variance σ^2. The DoG detector finds blob-like features, but also responds to edges. Keypoints on edges can be unstable, though, since their localization is not clearly defined.

Harris-Laplace Detector (Har-Lap). The Harris-Laplace detector [11] detects keypoints by a scale-adapted Harris function. It then selects the points for which the Laplacian-of-Gaussian reaches a maximum over scale. Similar to the Harris detector, this method finds corner-like keypoints, but in a scale-invariant fashion.

Hessian-Laplace Detector (Hes-Lap). The Hessian-Laplace detector [12] searches for local maxima of the Hessian determinant. As in the Harris-Laplace detector, a Laplacian-of-Gaussian is used for selecting maxima over scale. The detector finds blob-like structures. Its localization accuracy in scale-space is higher than for the DoG detector. Moreover, the scale selection accuracy is better than for Harris-Laplace (see [12]).

2.4 Local Descriptors

In this paper, we compare two shape-based local feature descriptors and apply them within the ISM framework.

Shape Context Descriptor. The Shape Context was originally developed to find matching points between object shapes [1]. Essentially, it builds for each point along a shape a log-polar histogram of edges around it. Edge structures close to the reference point are sampled in more detail than structures further away, as histogram bins become larger with increasing radius. This is a property equivalent to the idea of Geometric Blur [2]. We use 9 bins for the histogram and distinguish 4 edge orientations for each bin, which results in a 36-dimensional descriptor (see [12]). Feature similarity is measured by Euclidean distance.

Local Chamfer Descriptor. Chamfer distances are usually only used for entire object silhouettes, which is why we call the resulting approach *global* Chamfer matching. The Local Chamfer descriptor is an extension of this scheme which uses only local areas or sub-parts of the object shape or edge structure. Thus, matching can be done for subparts of the object. The edge structure is computed with the Canny detector. To obtain comparable results to those of [8], we use the same feature size of 25x25 pixels. Feature similarity is measured by the Chamfer distance between the edge structures. In order to obtain a codebook of Local Chamfer descriptors, we also use the Chamfer distance for clustering.
Local Chamfer Descriptor and Shape Context Descriptor using real edge images. Traditionally, Chamfer matching approaches, as well as Shape Context descriptors are trained on silhouette images. In the later test stage, however, they are applied to real edge images of which the silhouette images are only an idealized approximation. In order to make the descriptors more powerful and realistic, we therefore train both descriptors not only on silhouette images, but also on real edge images. When learning shape-based features on real edge images rather than on idealistic silhouette images, foreground as well as background structures influence the resulting features considerably. This is due to the fact that features may be localized only partly on the object boundary, and thus feature values can sometimes be dominated by background instead of foreground structures. As we will see in the experiments, however, this approach improves the detection results significantly, in particular for the Shape Context descriptor.

3 Experimental Evaluation

The aim of this paper is to evaluate the performance of various interest point detectors and different shape descriptors for pedestrian detection. The evaluation was conducted on two different training and test sets containing side views of pedestrians. Training set *A* and test set *A* contain images of people walking on the sidewalk or on the street. The images are recorded in front of two different backgrounds (the same for training and test set). In total, the training set consists of 105 images. The test set contains 197 images.

Image set *B* is more challenging. The images are recorded in common traffic scenes and pedestrian zones. Pedestrian appearances vary considerably, and people often take up only a small portion of the image. Also the backgrounds are far more varied and contain a significant amount of clutter. Figure 3 and 4 show some examples. From this data, we use 108 images as training set *B* and another set of 181 images as test set *B*. In particular, we made sure that none of the pedestrians, nor the image backgrounds occurred both in the training and the test set. For the results reported below, we use training set *A* together with training set *B* and evaluate the performance on test set *B*. We also used training set *B* only and obtained results similar to the ones reported below (they were omitted because of space constraints).

For both training sets, segmentation masks are available. These are either computed from the recorded image sequences with a Grimson-Stauffer background model [22] or manually annotated. Pedestrians in the test sets are annotated with bounding boxes.

For the evaluation, both test set *A* and *B* were rescaled, in order to enable a fair comparison with Harris points, which are only capable of single-scale detection. DoG, Harris-Laplace and Hessian-Laplace detectors are scale-invariant and therefore enable multi-scale detection.

Evaluation Criteria. The detection quality is measured by three criteria: *cover, overlap* and *relative distance*. *Cover* and *overlap* measure how much the annotation rectangle is covered by the detection hypothesis and vice versa. For the *relative distance*, we consider the distances between the bounding box centers of annotation and detection rectangle. We inscribe an ellipse in the annotation rectangle and relate the measured distance to the ellipse's radius at the corresponding angle. In the experiments only hypotheses with values of more than 50% for *cover* and *overlap* and less than 0.5 for the *relative distance* are accepted as correct.

3.1 Results on Test Set *A*

In a first step, we investigate the performance of the different interest point detectors for the Local Chamfer and Shape Context descriptors on test set *A*. Moreover, we explore how the use of real edge images during training influences detection performance.

Figures 5(a) and (b) show the performance of Local Chamfer and Shape Context trained on object silhouettes alone. With an equal error rate performance (EER) of up to 80% local Chamfer matching achieves better results than Shape Context with only around 63% EER. Figures 5(c) and (d) show the same detectors, but now trained on real edge images. In both cases we observe a clear improvement in performance, especially for Shape Context. Overall, Shape Context trained on real edge images (91% EER) outperforms all other combinations (up to 84% EER). This is an interesting result, in particular as it has been reported previously [23] that Shape Context matching does not perform well for cluttered scenes. From our experiments, one might even conclude that it is necessary to use real edge images to achieve good performance.

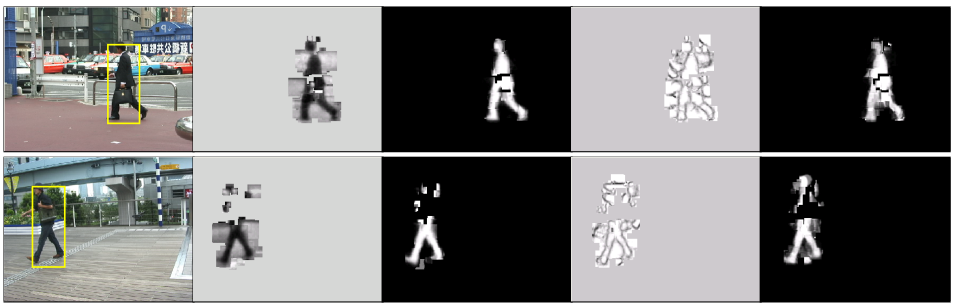
In the same Figures 5(a) through (d), one can observe great differences in detection performance depending on the choice of the interest point detector. The Hessian-Laplace detector performs best for most cases. Visually, it finds more points localized on pedestrians than the DoG detector and less on the background (see also Fig. 2). The Harris-Laplace detector also responds well on pedestrians, but detects less points in total. The Harris detector finds mainly corner points, which occur on both the pedestrian and the background.

We also compared the above results to global Chamfer matching and the method of [8], which is based on local image patches of size 25x25 instead of local shape descriptors. Figure 5(e) shows the respective results. The method of [8] achieves performances comparable to those for the local Chamfer matching. Global Chamfer matching with 4 orientation planes achieves only 77% EER performance, compared to 91% for the Shape Context descriptor. The local shape-based approaches therefore seem to be better suited for pedestrian detection than global Chamfer matching alone.

3.2 Results on Test Set *B*

As already mentioned, test set *B* is a more challenging image set. Nevertheless, the conclusions drawn from test set *A* still hold and can be reproduced on this test set as well.

With the new training and test set, we performed the same evaluation as before. We report only the details for the Shape Context descriptor, since again it performed considerably better than the other approaches. In combination with the Hessian-Laplace detector it achieves an EER performance of 89% (see Fig. 5(f)). The Harris-Laplace, DoG and



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

Figure 3: *Example detections on test set B with support images and segmentation masks ((b) support and (c) segmentation for Shape Context; (d) support and (e) segmentation for Local Chamfer).*

Harris detectors achieve 83%, 72% and 42% EER, respectively. This is consistent with the results for test set *A*, where we observed the same ranking for the different interest point detectors.

Integrated over the complete test set, the Hessian-Laplace detector finds about $130,000$ interest points; Harris and DoG around $100,000$; and Harris-Laplace $70,000$. One could argue that more interest points lead to better detection results. In order to verify this, we performed another experiment and increased the thresholds for the Hessian-Laplace detector such that it also yielded only around $100,000$ points. Indeed, we observed a drop in performance to 87% EER. However, this result still outperforms the other detectors, which suggests that the quality of the interest points is higher as well.

The Shape Context descriptor trained on real edge images again achieves better results than with plain object silhouettes (we only show the best results obtained with the Hessian-Laplace detector in Fig.5(g)). For test set *B*, this is more surprising than for test set *A*. This time, training and test backgrounds were different. Thus, the Shape Context descriptor trained on real edge images seems to generalize well even when training and test backgrounds vary substantially. Again the Hessian-Laplace detector outperforms all other detectors (89% EER). Additionally as can be seen in Figure 5(g), the Shape Context descriptor again performs better than all other local descriptors such as Local Chamfer and local image patches, which achieve equal error rates below 70%. Global Chamfer matching performs very poorly on test set *B* with an EER of 21%.

The results for test set *A* and test set *B* show that Shape Context descriptors trained on real edge images together with the Hessian-Laplace detector seem to be particularly well-suited for pedestrian detection. Figure 4 shows some example detections (row 1 and 2) and typical false positives for this approach (row 3). False positives are sometimes obtained on background structures with similar edge patterns, such as columns or lamp posts. In other failure cases, localization of the pedestrian is not precise enough, and the detection bounding box contains the pedestrian only partially. Quite interesting is the last example image, where a pedestrian in the background is detected which was just not annotated. As training set *B* contains only three pedestrians in frontal views, we did not expect to detect any non-side-view pedestrians in the test set.

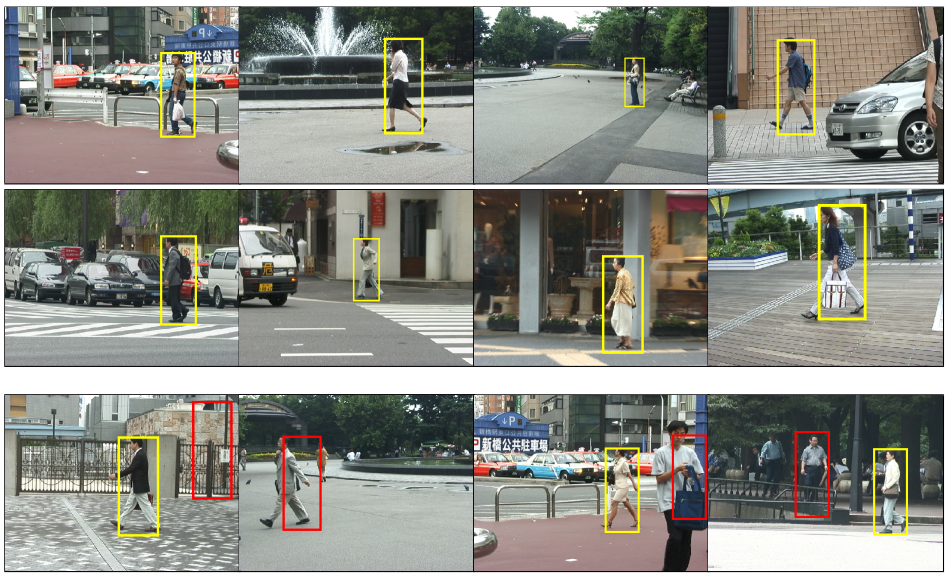


Figure 4: *Example detections at the EER on test set B drawn in yellow (row* 1 *and* 2*) and typical false positives drawn in red (row* 3*).*

4 Conclusion

The main aim of this paper was to compare various approaches for pedestrian detection. Quite interestingly various local approaches based on local Chamfer matching, Shape Context or image patches outperformed the standard technique of global Chamfer matching. Training the Local Chamfer and Shape Context descriptors on real edge images rather than on object silhouettes alone resulted in a substantial improvement in detection performance. Overall, the Shape Context descriptor trained on real edge images performed best, particularly on difficult images and backgrounds. Compared to raw image patches and Local Chamfer, it achieved a gain in EER performance of up to 20%.

The different interest point detectors had a large impact on detection performance, as well. The scale-invariant detectors perform better than the single-scale Harris detector. From our experience, this is due to the fact that codebooks from multi-scale detectors contain both small and large object structures, which can help to disambiguate detections. Overall the Hessian-Laplace detector outperformed the other detectors for all tested local descriptors.

In order to further confirm our results, we conducted an additional test on the data set used in [9]. This data set contains very challenging images with multiple overlapping pedestrians at various scales. As can be seen from Figure 5(h), Shape Context and Hessian-Laplace achieve a considerable improvement for this data set as well. The EER performance is slightly higher than the results reported in [9], where a supplementary Chamfer verification stage is applied. When a larger training set, as in section 3.2, is used, these results can be improved even further to a final EER of 83%.

In conclusion, using Shape Context descriptors trained on real edge images and the Hessian-Laplace detector represents a good combination for pedestrian detection. They

outperform the previously used combination of image patches and DoG [8, 9]. Moreover the Shape Context descriptor has a relatively low dimensionality, which speeds up the computation of the hypotheses.

In future work we plan to extend this study to other types of detectors, descriptors and also compare different recognition models. Moreover, we want to investigate the combination of various local features as well as the combination of local and global information in order to further improve detection performance.

Acknowledgments: This work has been funded, in part, by the EU project CoSy (IST-2002-004250)

References

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

[2] A. Berg and J. Malik. Geometric blur for template matching. In *CVPR*, pages 607–615, 2001.

[3] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *CVPR*, 2000.

[4] D. Forsyth and M. Fleck. Body plans. In *CVPR*, 1997.

[5] D. Gavrila. Pedestrian detection from a moving vehicle. In *ECCV*, pages 37–49. Springer, 2000.

[6] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.

[7] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV'04 Workshop on Statistical Learning in Computer Vision*, 2004.

[8] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC*, 2003.

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

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *IJCV*, 2004.

[11] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV*, 2001.

[12] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. Submitted to PAMI, 2004.

[13] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV*, pages 69–81, 2004.

[14] A. Mohan, C. Papageorgiu, and T. Poggio. Example-based object detection in images by componentes. In *PAMI*, pages 349–361, 2001.

[15] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *ECCV*, pages 666–680, 2002.

[16] G. Mori, X. Ren, A.A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR*, pages 326–333, 2004.

[17] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates, 1997.

[18] C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 2000.

[19] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *ECCV*, 2002.

[20] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *PAMI*, pages 530–535, 1997.

[21] A. Shashua, Y. Gdalyahu, and G. Hayon. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. In *IEEE Intelligent Vehicles Symposium*, Parma, Italy, 2004.

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

[23] A. Thayananthan, B. Stenger, P.H.S. Torr, and R.Cipolla. Shape context and chamfer matching in cluttered scenes. In *CVPR*, pages 127–135, Madison, Wisconsin, 2003.

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

[25] L. Zhao and C. Thorpe. Stereo and neural network-based pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 1(3):148 –154, 2000.

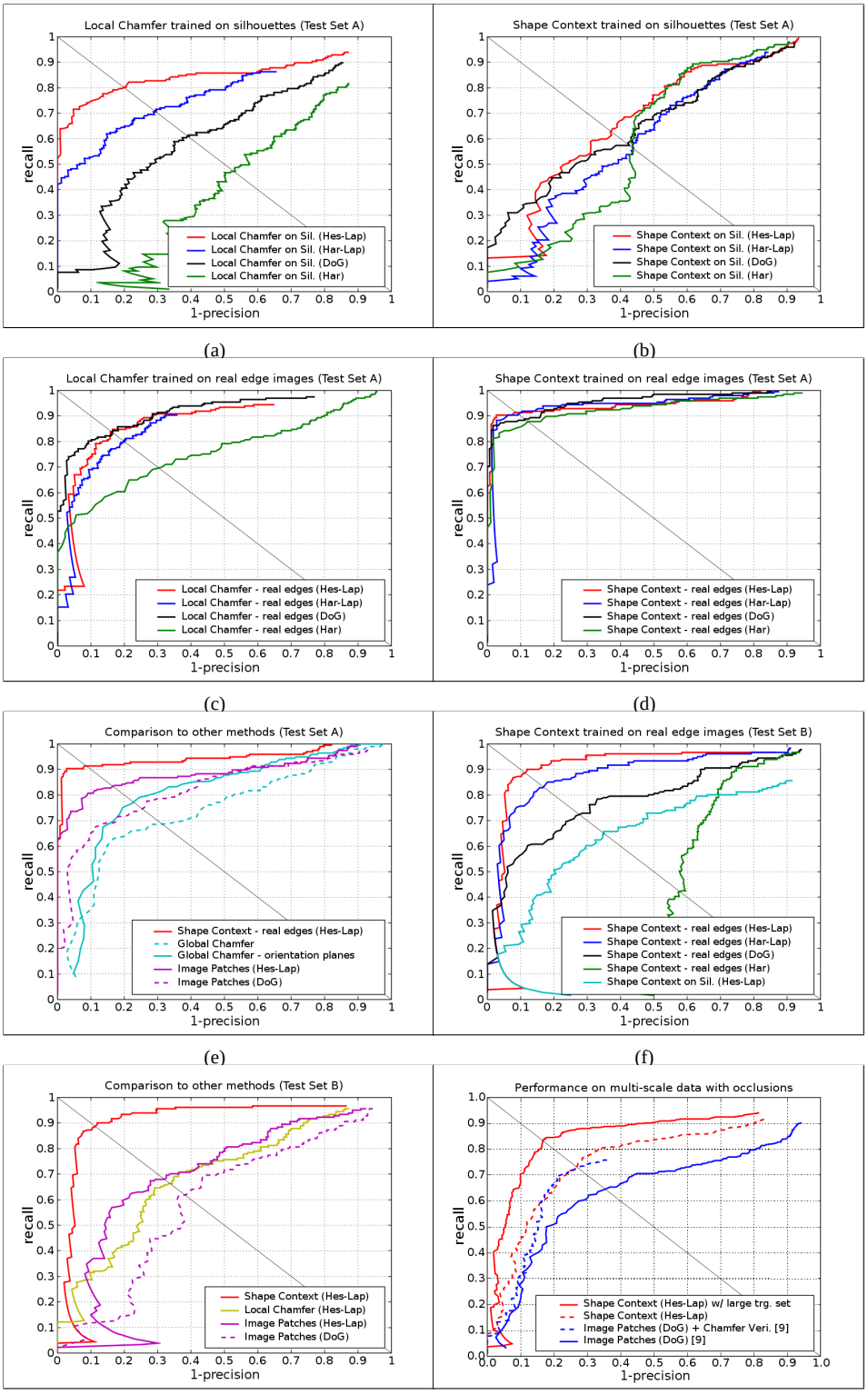


Figure 5: *Results on test set A (Plots (a)-(e)). Results on test set B (Plots (f) and (g)). Comparison to the results reported in [9] for multi-scale detection (Plot (h)).*