Pedestrian Detection in Crowded Scenes

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

Abstract

In this paper, we address the problem of detecting pedestrians in crowded real-world scenes with severe overlaps. Our basic premise is that this problem is too difficult for any type of model or feature alone. Instead, we present a novel algorithm that integrates evidence in multiple iterations and from different sources. The core part of our method is the combination of local and global cues via a probabilistic top-down segmentation. Altogether, this approach allows to examine and compare object hypotheses with high precision down to the pixel level. Qualitative and quantitative results on a large data set confirm that our method is able to reliably detect pedestrians in crowded scenes, even when they overlap and partially occlude each other. In addition, the flexible nature of our approach allows it to operate on very small training sets.

1. Introduction

The ability to reliably detect pedestrians in real-world images is interesting for a variety of applications, such as video surveillance or automatic driver-assistance systems in vehicles. At the same time, pedestrians are one of the most challenging categories for object detection. A large variability in their local and global appearance is caused by various types and styles of clothing, so that only few local regions are really characteristic for the entire category. In addition, the global shape undergoes a large range of transformations due to the variety of possible articulations and a multitude of occluding accessories such as backpacks, briefcases, and hand- or shopping bags, which may perturb a pedestrian's silhouette. Finally, in many applications several persons may be present in the same image region, partially occluding each other and adding to the difficulty.

Previous approaches to pedestrian detection have used either global models, e.g. using full-body appearance [20] or silhouettes [11, 10, 8]; or an assembly of local feature [23] or part detectors [18, 17]. However, only the latter two systems have been demonstrated under partial occlusion. As of today, no method has been evaluated for pedestrian detection in crowded scenes with strong overlaps, such as the ones in Fig. 1, despite the fact that these kinds of scenes often occur in real-world applications.

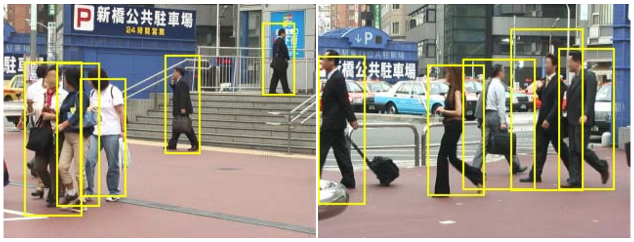
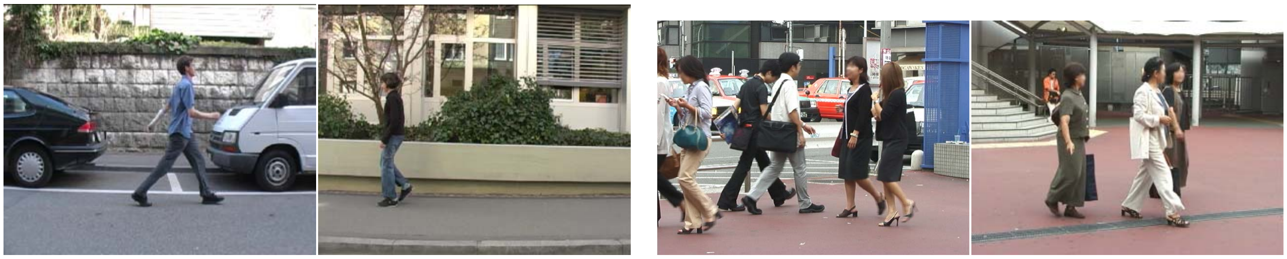


Figure 1. *Crowded real-world street scenes and our method's detections in them (shown in yellow).*

In this paper, we specifically address the task of detection in crowded scenes. The goal is to detect all pedestrians in a scene, localize them in the image and, if possible, infer their exact articulations. The difficulty of this task makes it problematic to rely on any type of model or feature alone. Instead, successful systems have to draw from the strengths of several approaches using appearance as well as shape cues and integrating both local and global information.

We follow this principle by aggregating evidence in several stages. We start by sampling local features from the image and combining them to generate hypotheses about possible object locations. For each hypothesis, we then compute a probabilistic top-down segmentation in order to determine its region of support in the image, which can be used to resolve ambiguities between overlapping hypotheses. As we will show, however, the additional difficulty of detecting pedestrians in crowded scenes makes it necessary to enforce also global constraints. For this goal, we propose a novel cue-integration scheme based on the hypothesized segmentations which facilitates this combination.

This paper contains the following contributions. 1) We present a new algorithm for evidence aggregation in multiple stages and from different sources. The foundation for this combination is an estimated object segmentation, which is used to integrate the influences of different cues on a common basis. 2) More specifically, we combine the local information from sampled appearance features with global cues about an object's silhouette. As our results demonstrate, this combination presents a way to make Chamfer matching robust to scale changes, clutter, and partial occlusion. 3) Our experiments show that the resulting system can reliably detect and localize pedestrians in difficult crowded scenes, even if they overlap and partially occlude each other.



(a) Training set

(b) Test set

Figure 2. *Example images from the training and test set for pedestrians used in our experiments*[1].

4) Last but not least, this performance is achieved using a training set that is between one and two orders of magnitude smaller than those of traditional approaches.

The paper is structured as follows. The following section defines the detection task we are going to address. Section 3 then presents our initial recognition method based on local appearance features. Motivated by the difficulty of the detection task, Section 4 proposes a novel cue integration scheme based on the hypothesized segmentation in the image. Finally, Section 5 presents experimental results.

2. Detection Task

The task we want to address is pedestrian detection in still gray images of crowded real-world scenes. For each pedestrian, we want to detect its presence, even when it is partially occluded, and precisely localize it in the image. Following the task set by [1], we also want to answer the question *how many* pedestrians are present in the scene. We therefore only accept a single hypothesis per pedestrian as correct and count each additional hypothesis on the same pedestrian as false positive.

2.1. Training Set

For our training data we recorded 44 sequences of 35 different people walking parallel to the camera image plane in front of two different backgrounds (see Fig. 2(a)). The training images include a wide range of different clothing and accessories, such as backpacks, hand bags, or books. As those images were taken in a Western city, the clothing styles invariably differ from those in the test set. Using the sequences as input, we computed a motion segmentation mask with a Grimson-Stauffer background model [21] and selected a set of frames for which a good segmentation could be obtained. For the local appearance-based approach described in Section 3, we used 105 training images and their reference segmentations; for the silhouette-based refinement stage in Section 4, we took 210 training images (plus in both cases their mirrored versions).

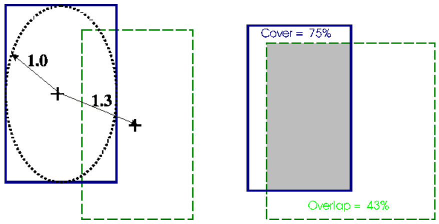


Figure 3. *Evaluation criteria for comparing bounding boxes: (left) relative distance; (right) cover and overlap.*

2.2. Test Set

Most current pedestrian detection systems have only been evaluated on images containing isolated pedestrians without large overlaps. It is our belief that in order to really improve the state of the art, research should proceed to more realistic and challenging scenarios. For this reason, we evaluate our algorithms on a highly difficult data set with crowded street scenes in an Asian metropolis. This test set consists of 209 images containing a total of 595 annotated pedestrians. Some examples[1] can be seen in Fig. 2(b).

The reason why we only speak of "annotated" pedestrians is that in the crowded scenes we are interested in, it is often not obvious where to draw the line and decide whether a pedestrian should be counted or not. In our test set, people occur in every state of occlusion, from fully visible to just half a leg protruding behind some other person. We therefore decided to annotate all those cases where a human could clearly detect the pedestrian without having to resort to reasoning. As a consequence, all pedestrians were annotated where at least some part of the torso was visible. However, this means that a good detector might still occasionally respond to pedestrians that are not annotated. On the other hand, a significant number of the annotated pedestrians are so severely occluded that it would be unrealistic to expect any current algorithm to achieve 100% recognition rate with just a small number of false positives.

2.3. Evaluation Criteria

For evaluating the detection results, we are not only concerned about a yes/no detection decision, but also about the pedestrians' precise locations and extents. We therefore

[1] For legal reasons, the pedestrians' faces shown in this paper had to be blurred. However, all experiments were performed on the original, unblurred images.

apply three criteria: *relative distance*, *cover*, and *overlap*. The *relative distance* d_r measures the distance between the bounding box centers in relation to the size of the annotation rectangle (see Fig. 3(left)). For this, we inscribe an ellipse in the annotation rectangle and relate the measured distance to the ellipse's radius at the corresponding angle. In this evaluation, the annotation rectangle had a fixed aspect ratio of 11:15. *Cover* and *overlap* measure how much of the annotation rectangle is covered by the detection hypothesis and vice versa (see Fig. 3(right)). Together, these criteria allow to compare hypotheses at different scales. In all following experiments, we consider a detection correct if $d_r \leq 0.5$ (corresponding to a deviation up to 25% of the true object size) and *cover* and *overlap* are both above 50%. As argued before, only one hypothesis per object is accepted as correct – any additional hypothesis on the same object is counted as a false positive.

3. Initial Recognition Approach

The first stage of our system generates object hypotheses by combining the evidence from local features [1, 9]. This stage is based on a scale-invariant extension of the Implicit Shape Model (ISM) [12, 14], which has been demonstrated to yield good recognition results for rigid object categories such as cars. Since our later steps closely build upon this approach, we will first briefly review its main components (see [13, 12, 14] for details).

3.1. Training

Training proceeds in two steps. We first learn a codebook of local appearances that are characteristic for the object category. This is done by applying a scale-invariant DoG interest point operator [16] to all training images and extracting image patches with a radius of 3σ of the detected scale. All extracted patches are then rescaled to a fixed size (in our case 25×25 pixels) and grouped using an agglomerative clustering scheme [12]. The resulting clusters form a compact representation of local object structure. Only the cluster centers are stored as codebook entries.

Next, we learn the spatial occurrence distribution of each codebook entry on the object category. For this, we perform a second iteration over all training images, again extracting patches around interest points, and record for each codebook entry all locations for which it could be matched to the extracted patches (where patch similarity is measured by *Normalized Greyscale Correlation* (NGC)). For each such occurrence, we additionally record the patch figure-ground map from the (segmented) training image. This information is used later in Section 3.3 to generate a top-down segmentation for each recognition hypothesis.

3.2. Initial Hypothesis Generation

During recognition, the same patch extraction procedure is applied, and the local information from sampled patches is collected in a probabilistic Hough voting procedure. Each patch is matched to the codebook, and matching codebook entries cast votes for possible object positions and scales according to their learned spatial probability distribution. In order to make this process robust, specific attention is paid to model the uncertainty on both levels: during the codebook matching stage and when inferring object locations. Formally, this is expressed as follows. Let $\mathbf{e}$ be an image patch observed at location ℓ. By matching it to the codebook, we generate probabilistic votes for different object categories o_n and locations $\lambda = (\lambda_x, \lambda_y, \lambda_s)$, which are weighted according to the following marginalization:

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

where $p(c_i|\mathbf{e})$ denotes the probability that patch $\mathbf{e}$ matches to codebook entry c_i, and $P(o_n, \lambda|c_i, \ell)$ describes the stored spatial probability distribution for the object center relative to an occurrence of that codebook entry. Object hypotheses are found as maxima in the 3D voting space using Mean Shift Mode Estimation [6] with a scale-adaptive *balloon density estimator*[2] [7] and a uniform cubical kernel K:

$$\hat{p}(o_n, \lambda) = \frac{1}{nh(\lambda)^d} \sum_k \sum_j p(o_n, \lambda_j|\mathbf{e}_k, \ell_k) K(\frac{\lambda - \lambda_j}{h(\lambda)}) \quad (2)$$

3.3. Top-down Segmentation

Recent approaches have demonstrated the possibility to generate top-down segmentations using learned knowledge about an object category [4, 24, 13]. Here, we use this idea to improve recognition. For each hypothesis, we go back to the image to determine on a per-pixel level where its support came from, thus effectively segmenting the object from the background. As shown in [13, 12], we can obtain the per-pixel probabilities of each pixel being *figure* or *ground* by the following double marginalization (first over patches then over codebook entries):

$$p(\mathbf{p} = \textit{figure}|o_n, \lambda) = \qquad (3)$$
$$\sum_{\mathbf{p} \in (\mathbf{e}, \ell)} \sum_i p(\mathbf{p} = \textit{fig.}|o_n, \lambda, c_i, \ell) \frac{p(o_n, \lambda|c_i, \ell) p(c_i|\mathbf{e}) p(\mathbf{e}, \ell)}{p(o_n, \lambda)}$$

where $p(\mathbf{p} = \textit{fig.}|o_n, \lambda, c_i, \ell)$ describes the local patch segmentation that can be inferred when codebook entry c_i is observed at location ℓ relative to the object center λ. The final segmentation is obtained by building the likelihood ratio between *figure* and *ground* probabilities. Figure 4 shows some of the resulting segmentations that can be obtained by this approach.

[2] For a discussion why this formulation is necessary, please see [14].

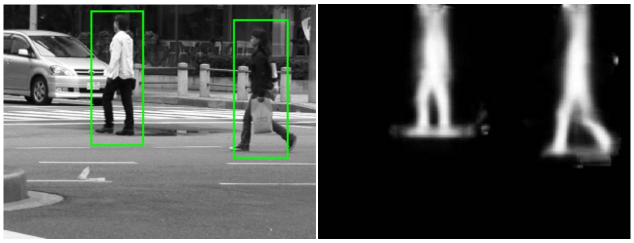


Figure 4. *Example recognition result and the corresponding top-down segmentation.*

3.4. Segmentation-based Verification

A central topic of this paper is to aggregate evidence from the image in several iterations. In [12], we have proposed a first step in this direction by using the top-down segmentation to refine object hypotheses in an MDL-based verification stage. The key idea behind this step is to integrate only information about the object itself and discard misleading influences from the background. At the same time, the segmentation reveals exactly from where in the image a hypothesis draws its support. Since each pixel can only be assigned to a single object, this step makes it possible to resolve ambiguities between overlapping hypotheses and search for the subset that best explains the image.

In this paper, we use an extended version of the MDL framework from [12]. Hypotheses are evaluated in terms of the *savings* [15] that can be obtained by explaining part of an image by the hypothesis h (see [12] for details):

$$S_h = K_0 S_{area} - K_1 S_{model} - K_2 S_{error} \tag{4}$$

In this formulation, S_{area} corresponds to the number N of pixels that can be explained by h; S_{error} denotes the cost for describing the error made by this explanation; and S_{model} describes the model complexity. Since objects at different scales take up different portions of the image, we make the model cost dependent on the *expected area* $A_s \sim \lambda_s^2$ an object occupies at a certain scale. As an estimate for the error we collect, over all pixels that belong to the segmentation of h, the probabilities that these pixels are not *figure*:

$$S_{error} = \sum_{\mathbf{p} \in Seg(h)} (1 - p(\mathbf{p} = \textit{figure}|h)). \tag{5}$$

Inserting (5) in (4) we obtain, after some simplifications,

$$S_h = -\frac{K_1}{K_0} + (1 - \frac{K_2}{K_0})\frac{N}{A_s} + \frac{K_2}{K_0}\frac{1}{A_s}\sum_{\mathbf{p} \in Seg(h)} p(\mathbf{p} = \textit{fig.}|h).$$

If multiple hypotheses h_1 and h_2 are present, we can derive the savings of the *combined hypothesis* $(h_1 \cup h_2)$:

$$S_{h_1 \cup h_2} = S_{h_1} + S_{h_2} - S_{area}(h_1 \cap h_2) + S_{error}(h_1 \cap h_2)$$

Based on this formulation, we search for the combination of hypotheses that maximize the savings using the method described in [12].

3.5. Experimental Results

Figure 7 shows the quantitative detection results on the test set in the form of a Recall-Precision Curve (RPC). It can be seen that the initial voting stage reaches a relatively low equal error rate (EER) performance of 27% (corresponding to 161 out of 595 correct detections with 434 false positives). This confirms the difficulty of the pedestrian detection task, compared to the detection of rigid object categories such as cars. However, the voting stage manages to find nearly all pedestrians eventually, which makes it well-suited as input to the next stage of our evidence aggregation procedure. As the figure shows, this segmentation-based verification loop presents a major improvement and raises the EER performance to 64%. In absolute numbers, this performance corresponds to 363 correct detections with 204 false positives. The improvement is emphasized even more by the fact that the initial voting stage returned 2,346 false positives for the same level of recall.

3.6. Discussion

An important factor to the method's performance is the flexibility of representation that is made possible by the Implicit Shape Model. As only consistency with a common object center is enforced, the approach can interpolate between local parts seen on different training objects. As a result, it only needs a relatively small number of training examples to recognize and segment categorical objects in different articulations and with widely varying texture patterns.

However, this flexibility is also an important restriction. Consider the examples in Figure 5. Here, the segmentations contain overlapping body parts from other pedestrians, such as another foot or a third leg. Each of those body parts is locally consistent with the object center – in the absence of other information, it could really belong to the object. As the ISM has no global knowledge about how many legs a pedestrian is supposed to have, these superfluous body parts are added to the hypothesis and may augment its recognition score on the expense of other hypotheses.

This effect is intrinsic to detection in crowded scenes. It would be even stronger if we did not refine hypotheses on the pixel level. However, also in our approach it has a negative impact on detection performance. Since the verification stage compares segmentations as a whole and assigns overlapping pixels to the highest-ranking candidate hypothesis, this effect may lead to important evidence being withheld from neighboring hypotheses and thus to lost detections. The following section therefore aims to alleviate this problem by adding global constraints.

4. Combination with Global Cues

As discussed above, the main restriction of the method presented so far is that the Implicit Shape Model, based only

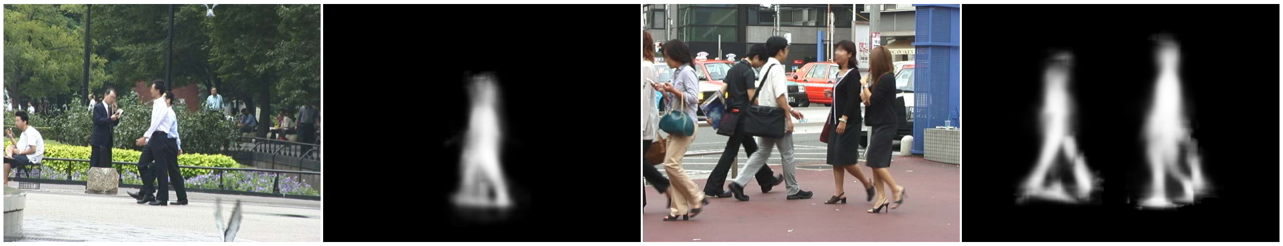


Figure 5. *Examples where the flexibility of the Implicit Shape Model leads to overcomplete segmentations. The images show the hypothesized segmentations for some of the pedestrians (only a subset of hypotheses is shown in order to reduce clutter). Since only consistency with a common object center is enforced, superfluous body parts may be assigned to a hypothesis. As a result of this wrong allocation, the scores of neighboring hypotheses may be reduced.*

on local features, has a very limited notion of global consistency. In the following, we want to enforce this consistency by adding the information from global shape cues.

4.1. Chamfer Matching

Chamfer matching was first proposed by [2] and later refined by [5, 19, 11] to detect objects based on global shape features. Given a set of trained shape templates (e.g. object silhouettes), Chamfer matching searches the image for locations where these templates can best be matched to the image content. Object shapes are compared using a distance transform (DT), which computes for each pixel the distance to the nearest feature pixel. Matches of a template T to the distance-transformed image I are found by shifting the template over the image and computing, at each location, the average distance value of all pixels that are covered by the template. The influence of outliers is reduced by using a truncated distance for matching:

$$D_{Chamfer}(T, I) = \frac{1}{|T|} \sum_{t \in T} \min(DT_I(t), \tau) \qquad (6)$$

The advantage of matching a template with the distance-transformed image rather than with the original edge image is that the resulting similarity measure will be smoother as a function of the template transformation parameters [10], which allows to speed up the matching process by employing a hierarchical coarse-to-fine search.

In this work, we want to use Chamfer matching in order to verify and refine object hypotheses by global constraints. As input, we take the set of hypotheses that is returned by the MDL verification stage from Section 3.4, operated at a permissive setting. Each hypothesis comes with a position and scale estimate and its probabilistic segmentation. In order to deal with different object scales, we use the scale estimate of the previous stage to cut out a surrounding region of the raw input image and rescale it such that the hypothesized object has uniform size (Fig. 6(a)). After that, we apply a Canny edge detector and compute the distance transform (Fig. 6(c)). For Chamfer matching, we use a set of 210 pedestrian silhouettes (plus their mirrored versions) extracted from the same training data as described in Section 2.1. As Chamfer matching is very sensitive to scale changes, we use 7 rescaled versions of each template, covering a scale range of [0.8,1.2].

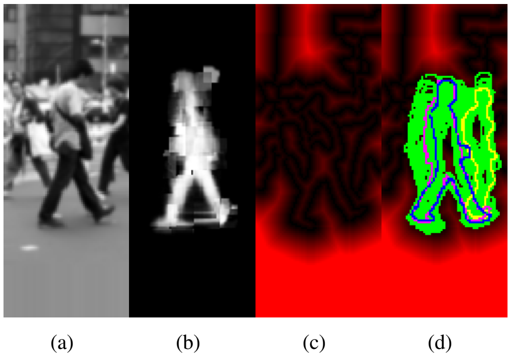


Figure 6. *Stages of the Chamfer verification procedure. (a) rescaled image region; (b) hypothesized segmentation; (c) distance-transformed image (d) fitted silhouettes (green: explored silhouettes, yellow: best Chamfer score, magenta: best fit to segmentation, blue: best combined fit).*

4.2. Combination with Segmentation

However, an important point we make is that Chamfer matching alone is not robust enough for the verification task. This can be visualized by the typical Chamfer matching results shown in Fig. 6(d). Although the correct contour is also found within the first 200 candidate matches, clutter and poor contrast cause the highest-ranking contour (shown in yellow) to lie on the background. As can be expected, this problem gets even worse the more shape templates and search scales are added in the matching process.

The reason for this is an intrinsic problem of Chamfer matching that occurs especially in cluttered images with many edges, where each pixel of the distance-transformed image contains only a relatively low value. As the Chamfer score is averaged over the whole silhouette, many different templates may reach approximately equal scores at various image locations and scales. In order to include the correct silhouette, one would therefore have to accept many false positives. The effect can be slightly alleviated by using multiple edge orientation planes, as proposed in [10, 22], but the general problem persists.

Especially rectangular structures are problematic, since they are, when averaged over the whole contour, very similar to certain pedestrian silhouettes. Unfortunately, such structures occur frequently in street scenes, e.g. on windows, columns, and lamp or fence posts. In addition, the spaces in-between pedestrians are often mistaken for additional hypotheses, since they exhibit similar edge patterns. However, those failure cases are complementary to those of our local approach. In the following, we therefore propose a way to combine the global shape information with the result of our local approach which makes Chamfer matching more robust to those situations.

The basis for our combination is the segmentation in the image. In particular, we search for a shape template that simultaneously maximizes the Chamfer score *and* the overlap with the hypothesized segmentation. The overlap can be expressed by the Bhattacharyya coefficient [3], which measures the affinity between two distributions. Assuming a uniform distribution for the points inside the shape template s, shifted to location q, we compare its overlap with the hypothesized segmentation Seg:

$$\rho(q) = \sum_x \sqrt{Seg(x)s(x,q)} \tag{7}$$

and compute a joint score as a linear combination

$$score = \alpha(1 - \frac{D_{Chamfer}}{\beta}) + (1-\alpha)\rho. \tag{8}$$

In our experiments, we set the weights to $\alpha = 0.45$ and $\beta = 50$, but we found the approach to be robust over a wide range of settings.

Fig. 6(d) shows the result of this combined estimation step. The resulting contour (shown in blue) integrates the information from local cues with a global fit to the image and provides better results than either method alone.

4.3. Combined MDL Verification

The above procedure is used to select a suitable silhouette $Silh(h)$ for each hypothesis h. We now adapt the verification procedure to use these global constraints and compute a hypothesis's savings not over the full segmentation anymore, but only over the part that is consistent with $Silh(h)$:

$$\tilde{S}_h = -\frac{K_1}{K_0} + (1 - \frac{K_2}{K_0})\frac{N}{A_s} + \frac{K_2}{K_0}\frac{1}{A_s}\sum_{\mathbf{p}\in Silh(h)} p(\mathbf{p} = \mathit{fig.}|h).$$

Moreover, we extend the verification procedure by formulating it as a quadratic Boolean optimization problem, similar to the one in [15]. Let $m^T = (m_1, m_2, \ldots, m_M)$ be a vector of indicator variables, where m_i has the value 1 if hypothesis h_i is present, and 0 if it is absent in the final description. In this formulation, the objective function for maximizing the savings takes the following form:

$$S(\hat{m}) = \max_m m^T Q m = m^T \begin{bmatrix} c_{11} & \cdots & c_{1M} \\ \vdots & \ddots & \vdots \\ c_{M1} & \cdots & c_{MM} \end{bmatrix} m. \tag{9}$$

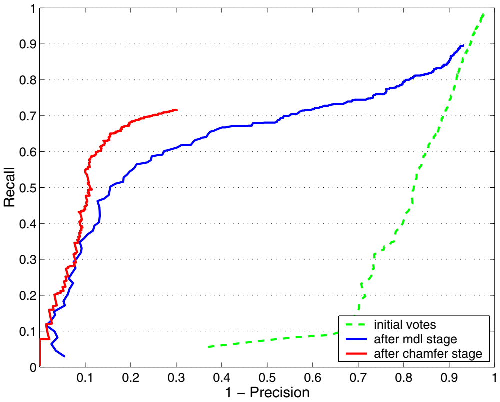


Figure 7. *Quantitative pedestrian detection results of our approach on the test set.*

The diagonal terms of Q express the savings of a particular hypothesis $c_{ii} = \tilde{S}_{h_i}$, while the off-diagonal terms handle the interaction between overlapping hypotheses (such that $\tilde{S}_{h_i \cup h_j} = c_{ii} + c_{jj} - 2c_{ij}$). Under the assumption that the stronger hypothesis opaquely occludes the weaker one, we obtain

$$c_{ij} = \frac{1}{2}\left(-(1 - \frac{K_2}{K_0})|O_{ij}| - \frac{K_2}{K_0}\sum_{\mathbf{p}\in O_{ij}} \min_{h_i,h_j} p(\mathbf{p} = \mathit{fig.}|h)\right)$$

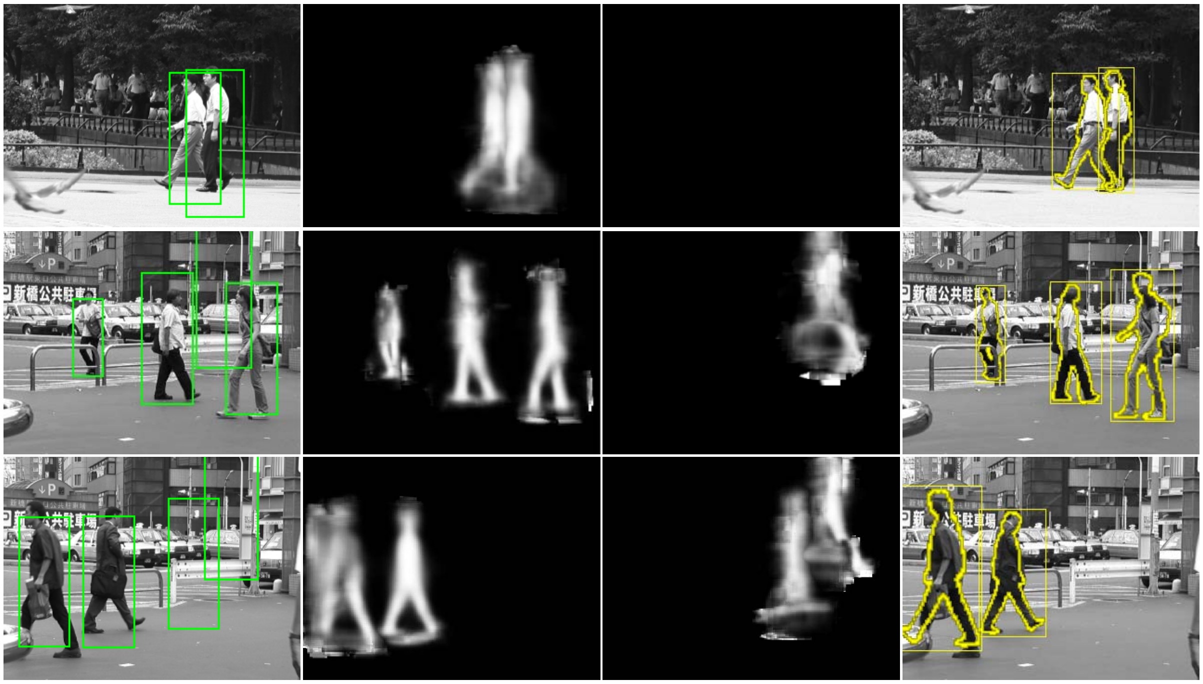
where $O_{ij} = Silh(h_i) \cap Silh(h_j)$ denotes the area of overlap between the (silhouette) segmentations of h_i and h_j.

Using this framework, the best image interpretation can be found using standard methods for quadratic boolean optimization problems. As the number of possible combinations grows exponentially with increasing problem size, though, it may become untractable to search for the globally optimal solution. In practice, we found it sufficient for our application to only compute a greedy approximation, as also argued in [15].

Fig. 8 shows some examples where this combined verification procedure is applied to difficult test images. As can be seen, the method is able to refine the correct hypotheses (e.g. Fig. 8(top)) and reject false positives on the background (Fig. 8(middle and bottom)). Note in particular the quality of the fitted silhouettes even when the pedestrians have low contrast to the background or are partially occluded.

5. Experimental Results

Fig. 7 shows the quantitative results when the verification stage is added to the system. As can be seen, the EER



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

Figure 8. *Examples for the Chamfer verification procedure. (a) initial hypotheses; (b) hypothesized segmentations for correct hypotheses; (c) segmentations for false positives; (d) fitted silhouettes.*

performance is considerably improved from 64% to 71.3%, corresponding to 424 out of 595 correct detections with only 171 false positives. Without the final verification stage, our system could reach this level of recall only at a reduced precision of 42.7%. This means that at the same recall rate, the Chamfer verification manages to reject 399 additional false positives. In addition, the quality of the obtained bounding boxes is in many cases significantly improved, which is however not quantified by our evaluation criterion.

As already discussed in Section 2.2, the quantitative results should be regarded with special consideration. Many annotated pedestrians are severely occluded, and the detection task is so difficult that a performance in the upper 90% range is far beyond the state of current computer vision systems. In order to give a better impression of our method's performance, Fig. 9 therefore shows obtained detection results on example images from the test set (at the EER). As can be seen from those examples, the presented method can reliably detect and localize pedestrians in crowded scenes and with severe overlaps.

6. Conclusion

In this paper, we have presented a novel algorithm for pedestrian detection in crowded scenes. Our method does not procede in one single run, but through a series of iterative evidence aggregation steps. Throughout this process, we integrate local and global cues via an automatically computed top-down segmentation. Altogether, this iterative approach allows us to examine and disambiguate between object hypotheses at a high level of precision. Qualitative and quantitative results show our method's capability to reliably detect pedestrians in crowded street scenes and with severe overlaps. Finally, the flexible nature of our approach permits our method to work with very small training sets and generalize to novel scenarios.

It is also important to emphasize that our method operates on still images. Although video sequences are used for training in order to avoid manual segmentation, we are not using any temporal continuity information for recognition. However, our system's capability to robustly detect pedestrians in crowded scenes and estimate their articulations shows an interesting potential also for use in object tracking applications, which we will explore in future work.

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

References

[1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV'02*, pages 113–130, 2002.

[2] H.G. Barrow, J.M. Tenenbaum, R.C. Bolles, and H.C. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *IJCAI'77*, 1977.

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

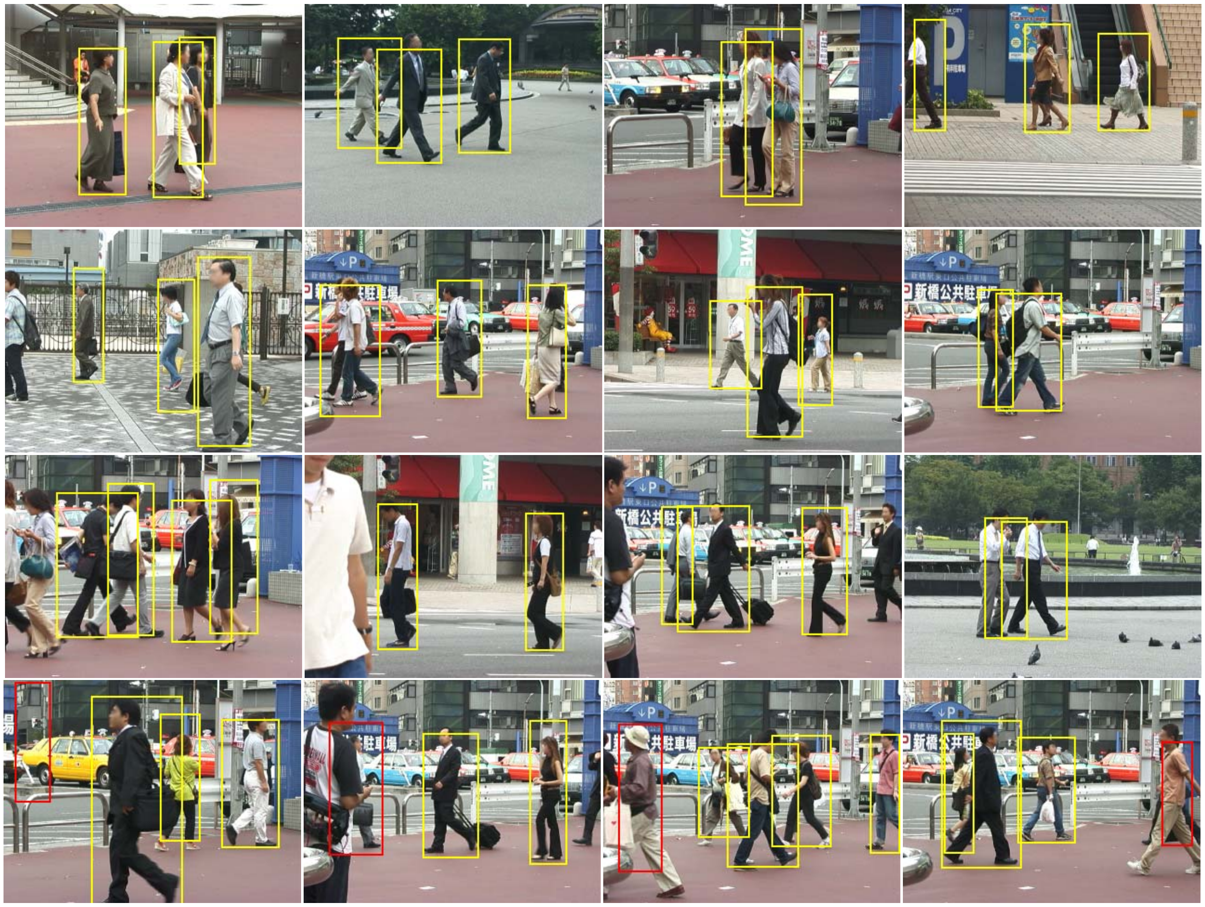


Figure 9. *Example detections of our approach on difficult crowded scenes from the test set (at the EER). Correct detections are shown in yellow, false positives in red. (bottom row): Examples for false positives. (left) true false positive; (middle left): correct detection, but not annotated; (middle right and right): bounding boxes estimated too small.*

[4] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV'02*, LNCS 2353, pages 109–122, 2002.

[5] G. Borgefors. Hierarchical chamfer matching: A parametric edge matching algorithm. *PAMI*, 10(6):849–865, 1988.

[6] D. Comaniciu and P. Meer. Distribution free decomposition of multivariate data. *Pattern Analysis and Applications*, 2(1):22–30, 1999.

[7] D. Comaniciu, V. Ramesh, and P. Meer. The variable bandwidth mean shift and data-driven scale selection. In *ICCV'01*, 2001.

[8] P. Felzenszwalb. Learning models for object recognition. In *CVPR'01*, 2001.

[9] R. Fergus, A. Zisserman, and P. Perona. Object class recognition by unsupervised scale-invariant learning. In *CVPR'03*, 2003.

[10] D. Gavrila. Pedestrian detection from a moving vehicle. In *ECCV'00*.

[11] D. Gavrila. Multi-feature hierarchical template matching using distance transforms. In *ICPR'98*, 1998.

[12] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV'04 Workshop on Stat. Learn. in Comp. Vis.*, pages 17–32, 2004.

[13] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC'03*, pages 759–768, 2003.

[14] 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, 2004.

[15] A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation of range images as the search for geometric parametric models. *IJCV*, 14:253–277, 1995.

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

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

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

[19] C.F. Olson and D.P. Huttenlocher. Automatic target recognition by matching oriented edge pixels. *Trans. Image Proc.*, 6(1):103–113, 1997.

[20] C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 38(1):15–33, 2000.

[21] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for realtime tracking. In *CVPR'99*, 1999.

[22] A. Thayananthan, B. Stenger, P.H.S. Torr, and R. Cipolla. Shape context and chamfer matching in cluttered scenes. In *CVPR'03*, 2003.

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

[24] S.X. Yu and J. Shi. Object-specific figure-ground segregation. In *CVPR'03*, 2003.