in *ECCV'04 Workshop on Statistical Learning in Computer Vision*, Prague, May 2004.

Combined Object Categorization and Segmentation with an Implicit Shape Model

Bastian Leibe[1], Ales Leonardis[2], and Bernt Schiele[1]

[1] Perceptual Computing and Computer Vision Group,
ETH Zurich, Switzerland
{leibe,schiele}@inf.ethz.ch
http://www.vision.ethz.ch/pccv

[2] University of Ljubljana, Faculty of Computer and Information Science,
SI-1001 Ljubljana, Slovenia
alesl@fri.uni-lj.si

Abstract. We present a method for object categorization in real-world scenes. Following a common consensus in the field, we do not assume that a figure-ground segmentation is available prior to recognition. However, in contrast to most standard approaches for object class recognition, our approach automatically segments the object as a result of the categorization.
This combination of recognition and segmentation into one process is made possible by our use of an Implicit Shape Model, which integrates both into a common probabilistic framework. In addition to the recognition and segmentation result, it also generates a per-pixel confidence measure specifying the area that supports a hypothesis and how much it can be trusted. We use this confidence to derive a natural extension of the approach to handle multiple objects in a scene and resolve ambiguities between overlapping hypotheses with a novel MDL-based criterion.
In addition, we present an extensive evaluation of our method on a standard dataset for car detection and compare its performance to existing methods from the literature. Our results show that the proposed method significantly outperforms previously published methods while needing one order of magnitude less training examples. Finally, we present results for articulated objects, which show that the proposed method can categorize and segment unfamiliar objects in different articulations and with widely varying texture patterns, even under significant partial occlusion.

将两者整合到一个共同的概率框架中。除了识别和分割的结果，它还生成了一个每个像素的置信度，用来指定支持假设的区域和它的可信程度。我们利用这种置信度推导出一种处理场景中多个物体的方法的自然扩展，并用一种新的基于mdl的标准解决重叠假设之间的歧义。

能够分类和分割不熟悉的物体在不同的关节和具有广泛变化的纹理模式，甚至在显著的部分遮挡下。

1 Introduction

The goal of our work is object categorization in real-world scenes. That is, given some training examples of an object category, we want to recognize a-priori unknown instances of that category and assign the correct category label. In order to transfer this capability to new domains, it is especially important that class characteristics be learned instead of hard-coded into the system. Therefore, we aim to learn solely from example images.

We pursue a two-staged approach. In the first step, we learn a *Codebook of Local Appearance* that contains information which local structures may appear on objects of

the target category. Next, we learn an *Implicit Shape Model* that specifies where on the object the codebook entries may occur. As the name already suggests, we do not try to define an explicit model for all possible shapes a class object may take, but instead define "allowed" shapes implicitly in terms of which local appearances are consistent with each other. The advantages of this approach are its greater flexibility and the smaller number of training examples it needs to see in order to learn possible object shapes. For example, when learning to categorize articulated objects such as cows, as described in Section 6, our method does not need to see every possible articulation in the training set. It can combine the information of a front leg seen on one training cow with the information of a rear leg from a different cow to recognize a test image with a novel articulation, since both leg positions are consistent with the same object hypothesis.

This idea is similar in spirit to approaches that represent novel objects by a combination of class prototypes [12], or of familiar object views [22]. However, the main difference of our approach is that here the combination does not occur between entire exemplar objects, but through the use of local image patches, which again allows a greater flexibility. Also, the Implicit Shape Model is formulated in a probabilistic framework that allows us to obtain a category-specific segmentation as a result of the recognition process. This segmentation can then in turn be used to improve the recognition results. In particular, we obtain a per-pixel confidence measure specifying how much both the recognition and the segmentation result can be trusted.

In [13], we describe an early version of this approach. However, this earlier paper contains only limited experimental evaluation, and the approach is restricted to scenes containing only one object. In this paper, we extend the method to handle multiple objects in a scene, effectively resolving ambiguities between overlapping hypotheses by a novel criterion based on the MDL principle. We also extensively evaluate the method on two large data sets and compare its performance to existing methods from the literature. Our results show a significant improvement over previously published methods. Finally, we present results for articulated objects, which show that the proposed method can categorize and segment unfamiliar objects in different articulations and with widely varying texture patterns. In addition, it can cope with significant partial occlusion.

The paper is structured as follows. The next section discusses related work. After that, we describe the recognition approach and its extension to generate category-specific segmentations. Section 4 then presents an evaluation on a car detection task. Using the segmentation obtained in the previous step, Section 5 extends the approach to resolve ambiguities between multiple object hypotheses with an MDL-based criterion and compares our performance to existing methods. Finally, Section 6 shows experimental results for the recognition and segmentation of articulated objects. A final discussion concludes our work.

2 Related Work

Various shape models have been used for the recognition of object classes. When regularly textured objects are used, the shape can be modelled by spatial frequency statistics of texture descriptors [20]. For detection and recognition of more general object classes, many current methods learn global or local features in fixed configurations [21, 19, 23].

Since they treat the object as a whole, such approaches need a large number of training examples. Others learn the assembly of hand-selected object parts using configuration classifiers [18] or by modelling the joint spatial probability distribution [4]. Weber & Perona [24] also learn the local parts and explicitly compute their joint distribution. Fergus et al. [9] extend this approach to scale-invariant object parts and estimate their joint spatial and appearance distribution. However, the complexity of this combined estimation step restricts their methods to a small number of parts. Agarwal & Roth [1] keep a larger number of object parts and apply a feature-efficient classifier for learning spatial configurations between pairs of parts. However, their learning approach relies on the repeated observation of cooccurrences between the same parts in similar spatial relations, which again requires a large number of training examples.

The idea to use top-down knowledge to drive the segmentation process has recently developed into an area of active research. Approaches, such as Deformable Templates [26], or Active Appearance Models [7], are typically used when the object of interest is known to be present in the image and an initial estimate of its size and location can be obtained. Borenstein & Ullman [3] generate class-specific segmentations by combining object fragments in a jigsaw-puzzle fashion. However, their approach is not integrated with a recognition process. Yu & Shi [25] present a parallel segmentation and recognition system in a graph theoretic framework, but only for a set of known objects.

Our approach integrates the two processes of recognition and segmentation in a common probabilistic framework. Given a set of training examples from an object class, it is able to automatically learn a category representation and recognize and segment a-priori unknown objects of this class in novel settings. By representing allowed part configurations in terms of an implicit model, it retains high flexibility while making efficient use of the available training data. The following sections describe how this combination is achieved.

3 Approach

An Implicit Shape Model $ISM(C) = (I_C, P_{I,C})$ for a given object category C consists of a class-specific alphabet I_C (in the following termed a *codebook*) of local appearances that are prototypical for the object category, and of a spatial probability distribution $P_{I,C}$ which specifies where each codebook entry may be found on the object.

In our definition, we impose two requirements for the probability distribution $P_{I,C}$. The first is that the distribution is defined independently for each codebook entry. This makes the approach flexible, since it allows to combine object parts during recognition that were initially observed on different training examples. In addition, it enables us to learn recognition models from relatively small training sets, as our experiments in Sections 4 and 6 demonstrate. The second constraint is that the spatial probability distribution for each codebook entry is estimated in a non-parametric manner. The method is thus able to model the true distribution in as much detail as the training data permits instead of making an oversimplifying Gaussian assumption.

The rest of this section explains how this learning and modeling step is implemented and how the resulting implicit model is used for recognition and segmentation.

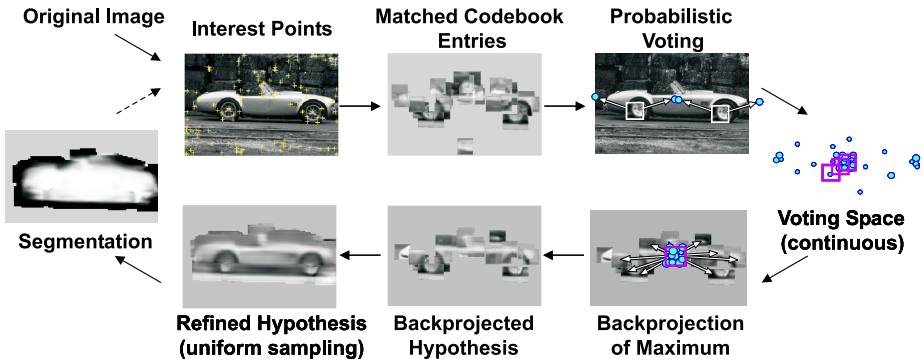


Fig. 1. *The recognition procedure. Image patches are extracted around interest points and compared to the codebook. Matching patches then cast probabilistic votes, which lead to object hypotheses that can later be refined. Based on the refined hypotheses, we compute a category-specific segmentation.*

3.1 A Codebook of Local Appearance

In order to generate a codebook of local appearances of a particular object category, we use an approach inspired by the work of Agarwal and Roth [1]. From a variety of images, patches of size 25×25 pixels are extracted with the Harris interest point detector [11]. Starting with each patch as a separate cluster, agglomerative clustering is performed: the two most similar clusters C_1 and C_2 are merged as long as the average similarity between their constituent patches (and thus the cluster compactness) stays above a certain threshold t:

$$similarity(C_1, C_2) = \frac{\sum_{p \in C_1, q \in C_2} NGC(p,q)}{|C_1| \times |C_2|} > t, \tag{1}$$

where the similarity between two patches is measured by Normalized Greyscale Correlation (NGC):

$$NGC(p,q) = \frac{\sum_i (p_i - \overline{p_i})(q_i - \overline{q_i})}{\sqrt{\sum_i (p_i - \overline{p_i})^2 \sum_i (q_i - \overline{q_i})^2}} \tag{2}$$

This clustering scheme guarantees that only those patches are grouped which are visually similar, and that the resulting clusters stay compact, a property that is essential for later processing stages. From each resulting cluster, we compute the cluster center and store it in the codebook.

Rather than to use this codebook directly to train a classifier, as in [1], we use them to define our Implicit Shape Model. For this, we perform a second iteration over all training images and match the codebook entries to the images using the NGC measure. Instead of taking the best-matching codebook entry only, we activate all entries whose similarity is above t, the threshold already used during clustering. For every codebook entry, we store all positions it was activated in, relative to the object center.

During recognition, we use this information to perform a Generalized Hough Transform [2, 15]. Given a test image, we extract image patches and match them to the codebook to activate codebook entries. Each activated entry then casts votes for possible positions of the object center. Figure 1 illustrates this procedure. It is important to emphasize that we use a continuous voting space in order to avoid discretization artefacts. We search for hypotheses as maxima in the continous voting space using Mean-Shift Mode Estimation [5, 6]. For promising hypotheses, all patches that contributed to it can be collected (Fig. 1(bottom)), therefore visualizing what the system reacts to. Moreover, we can refine the hypothesis by sampling all the image patches in its surroundings, not just those locations returned by the interest point detector. As a result, we get a representation of the object including a certain border area.

3.2 Probabilistic Formulation

In the following, we cast this recognition procedure into a probabilistic framework (extending the framework from [13]). Let **e** be our evidence, an extracted image patch observed at location ℓ. By matching it to our codebook, we obtain a set of valid interpretations I_i. Each interpretation is weighted with the probability $p(I_i|\mathbf{e}, \ell)$. If a codebook cluster matches, it can cast its votes for different object positions. That is, for every I_i, we can obtain votes for several object identities o_n and positions x, which we weight with $p(o_n, x|I_i, \ell)$. Formally, this can be expressed by the following marginalization:

$$p(o_n, x|\mathbf{e}, \ell) = \sum_i p(o_n, x|\mathbf{e}, I_i, \ell) p(I_i|\mathbf{e}, \ell). \tag{3}$$

Since we have replaced the unknown image patch by a known interpretation, the first term can be treated as independent from **e**. In addition, we match patches to the codebook independent of their location. The equation thus reduces to

$$p(o_n, x|\mathbf{e}, \ell) = \sum_i p(o_n, x|I_i, \ell) p(I_i|\mathbf{e}). \tag{4}$$

$$= \sum_i p(x|o_n, I_i, \ell) p(o_n|I_i, \ell) p(I_i|\mathbf{e}). \tag{5}$$

The first term is the probabilistic Hough vote for an object position given its identity and the patch interpretation. The second term specifies a confidence that the codebook cluster is really matched on the object as opposed to the background. This can be used to include negative examples in the training. Finally, the third term reflects the quality of the match between image patch and codebook cluster.

By basing the decision on single-patch votes and assuming a uniform prior for the patches, we obtain

$$score(o_n, x) = \sum_k \sum_{x_j \in W(x)} p(o_n, x_j|\mathbf{e}_k, \ell_k). \tag{6}$$

From this probabilistic framework, it immediately follows that the $p(I_i|\mathbf{e})$ and $p(x|o_n, I_i, \ell)$ should both sum to one. In our experiments, we assume a uniform distribution for both

(meaning that we set $p(I_i|\mathbf{e}) = \frac{1}{|I|}$, with $|I|$ the number of matching codebook entries), but it would also be possible, for example, to let the $p(I_i|\mathbf{e})$ distribution reflect the relative matching scores.

By this derivation, we have embedded the Hough voting strategy in a probabilistic framework. In this context, the mean-shift search over the voting space can be interpreted as a Parzen window probability density estimation for the correct object location. The power of this approach lies in its non-parametric nature. Instead of making Gaussian assumptions for the codebook cluster distribution on the object, our approach is able to model the true distribution in as much detail as is possible from the observed training examples. 通过这种推导，我们将霍夫投票策略嵌入到一个概率框架中。在这种情况下，投票空间上的均值偏移搜索可以解释为正确目标位置的Parzen窗概率密度估计。这种方法的强大之处在于它的非参数特性。相反的假设目标上的码本聚类分布为高斯分布，我们的方法能够从观察到的训练例子中尽可能详细地模拟真实分布。

3.3 Object Segmentation

In this section, we describe a probabilistic formulation for the segmentation problem (as derived in [13]). As a starting point, we take a refined object hypothesis $h = (o_n, x)$ obtained by the algorithm from the previous section. Based on this hypothesis, we want to segment the object from the background.

Up to now, we have only dealt with image patches. For the segmentation, we now want to know whether a certain image pixel $\mathbf{p}$ is *figure* or *ground*, given the object hypothesis. More precisely, we are interested in the probability $p(\mathbf{p} = \textit{figure}|o_n, x)$. The influence of a given patch $\mathbf{e}$ on the object hypothesis can be expressed as

$$p(\mathbf{e}, \ell|o_n, x) = \frac{p(o_n, x|\mathbf{e}, \ell)p(\mathbf{e}, \ell)}{p(o_n, x)} = \frac{\sum_I p(o_n, x|I, \ell)p(I|\mathbf{e})p(\mathbf{e}, \ell)}{p(o_n, x)} \tag{7}$$

where the patch votes $p(o_n, x|\mathbf{e}, \ell)$ are obtained from the codebook, as described in the previous section. Given these probabilities, we can obtain information about a specific pixel by marginalizing over all patches that contain this pixel:

$$p(\mathbf{p} = \textit{figure}|o_n, x) = \sum_{\mathbf{p}\in(\mathbf{e},\ell)} p(\mathbf{p} = \textit{figure}|o_n, x, \mathbf{e}, \ell)p(\mathbf{e}, \ell|o_n, x) \tag{8}$$

with $p(\mathbf{p} = \textit{figure}|o_n, x, \mathbf{e}, \ell)$ denoting patch-specific segmentation information, which is weighted by the influence $p(\mathbf{e}, \ell|o_n, x)$ the patch has on the object hypothesis. Again, we can resolve patches by resorting to learned patch interpretations I stored in the codebook:

$$p(\mathbf{p} = \textit{figure}|o_n, x) = \sum_{\mathbf{p}\in(\mathbf{e},\ell)} \sum_I p(\mathbf{p}=\textit{fig.}|o_n, x, \mathbf{e}, I, \ell)p(\mathbf{e}, I, \ell|o_n, x) \tag{9}$$

$$= \sum_{\mathbf{p}\in(\mathbf{e},\ell)} \sum_I p(\mathbf{p}=\textit{fig.}|o_n, x, I, \ell)\frac{p(o_n, x|I, \ell)p(I|\mathbf{e})p(\mathbf{e}, \ell)}{p(o_n, x)} \tag{10}$$

This means that for every pixel, we build a weighted average over all segmentations stemming from patches containing that pixel. The weights correspond to the patches' respective contributions to the object hypothesis. For the *ground* probability, the result is obtained in an analogue fashion. 这意味着对每一个像素建立了一个加权平均，所有的分割来自于包含该像素的patch。权重对应于patch各自对目标假设的贡献。对于ground truth概率，采用模拟方法得到结果。

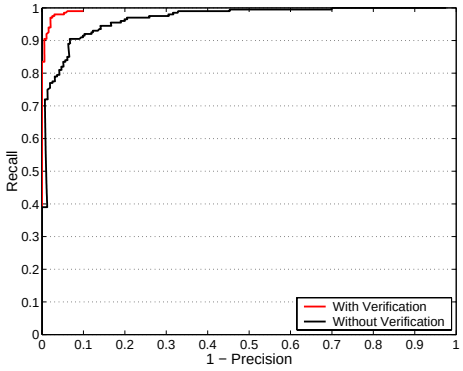


Fig. 2. *Results on the UIUC car database with and without the MDL hypothesis verification stage.*

The most important part in this formulation is the per-pixel segmentation information $p(\mathbf{p} = \textit{figure}|o_n, x, I, \ell)$, which is only dependent on the matched codebook entry, no longer on the image patch. In our approach, we implement this probability by keeping a separate segmentation mask for every stored *occurrence position* of each codebook entry. These patch figure-ground masks are extracted from a reference segmentation given for each training image. Further, we assume uniform priors for $p(\mathbf{e}, \ell)$ and $p(o_n, x)$, so that these elements can be factored out of the equations. In order to obtain a segmentation of the whole image from the figure and ground probabilities, we build the likelihood ratio for every pixel:

$$L = \frac{p(\mathbf{p} = \textit{figure}|o_n, x)}{p(\mathbf{p} = \textit{ground}|o_n, x)}. \tag{11}$$

Figure 7 shows example segmentations of cars, together with $p(\mathbf{p} = \textit{figure}|o_n, x)$, the system's confidence in the segmentation result. The darker a pixel, the higher its probability of being *figure*. The lighter it is, the higher its probability of being *ground*. The uniform gray region in the background of the segmentation image does not contribute to the object hypothesis and is therefore considered neutral. The estimate of how much the obtained segmentation can be trusted is especially important when the results shall later be combined with other cues for recognition or segmentation. It is also the basis for our MDL-based hypothesis selection criterion described in Section 5.

4 Results

In the early version presented in [13], our method has only been evaluated on small datasets. In the rest of this paper, we therefore present an extensive evaluation on two large databases, as well as a novel hypothesis verification stage based on the MDL criterion, which resolves ambiguities between overlapping hypotheses and handles scenes containing multiple objects

In order to compare our method's performance to state-of-the-art approaches, we applied it to the UIUC car database [1]. This test set consists of 170 images containing

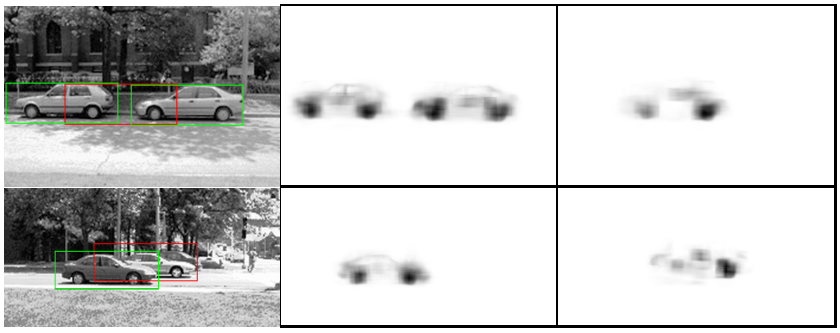


Fig. 3. *(left) Two examples for overlapping hypotheses (in red); (middle)* $p(\boldsymbol{p} = \textit{figure}|h)$ *probabilities for the front and (right) for the overlapping hypotheses. As can be seen, the overlapping hypothesis in the above example is fully explained by the two correct detections, while the one in the lower example obtains additional support from a different region in the image.*

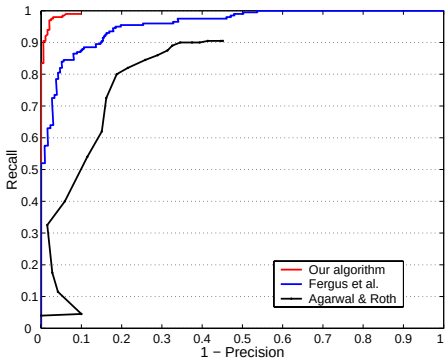
a total of 200 sideviews of cars. The images include instances of partially occluded cars, cars that have low contrast with the background, and images with highly textured backgrounds. In the dataset, all cars are approximately the same size.

Together with the test set, Agarwal & Roth provide a training set of 550 car and 500 non-car images. In our experiments, we do not use this training set, but instead train on a much smaller set of only 50 hand-segmented images (mirrored to represent both car directions) that were originally prepared for a different experiment. In particular, our training set contains both European and American cars, whereas the test set mainly consists of American-style sedans and limousines. Thus, our detector remains more general and is not tuned to the specific test conditions. The original data set is at a relatively low resolution (with cars of size 100*40 pixels). Since our detector is learned at a higher resolution, we rescaled all images by a constant factor prior to recognition (Note that this step does not increase the images' information content). All experiments were done using the evaluation scheme and detection tolerances from [1].

Figure 2 shows a recall-precision curve (RPC) of our method's performance. The plot was generated using the evaluation scheme and the detection tolerances from [1]. As can be seen from the figure, our method succeeds to generalize from the small training set and achieves an excellent performance with an Equal Error Rate (EER) of 91%. Analyzing the results on the test set, we observed that a large percentage of the remaining false positives are due to *secondary hypotheses*, which contain only one of the car's wheels, e.g. the rear wheel, but hypothesize it to be the front wheel of an adjoining car (see Figure 3 for an example). This problem is particularly prominent in scenes that contain multiple objects. The following section derives a hypothesis verification criterion which resolves these ambiguities in a natural fashion and thus improves the recognition results.

5 Multiple-Object Scene Analysis

As already mentioned in the previous section, a large number of the initial false positives are due to secondary hypotheses which overlap part of the object. This is a com-



Method	Equal Error Rate
Agarwal & Roth [1]	∼79%
Garg et al. [10]	∼88%
Fergus et al. [9]	88.5%
Our algorithm	97.5%

Fig. 4. *Comparison of our results on the UIUC car database with others reported in the literature.*

mon problem in object detection. Generating such hypotheses is a desired property of a recognition algorithm, since it allows the method to cope with partial occlusions. However, if enough support is present in the image, the secondary detections should be sacrificed in favor of other hypotheses that better explain the image. Usually, this problem is solved by introducing a bounding box criterion and rejecting weaker hypotheses based on their overlap. However, such an approach may lead to missed detections, as the example in Figure 3 shows. Here the overlapping hypothesis really corresponds to a second car, which would be rejected by the simple bounding box criterion (Incidentally, only the front car is labeled as "car" in the test set, possibly for exactly that reason). However, since our algorithm provides us with an object segmentation together with the hypotheses, we can improve on this. In the following, we derive a criterion based on the principle of Minimal Description Length (MDL), inspired by the approach pursued in [14].

The MDL principle is an information theoretic formalization of the general notion to prefer simple explanations to more complicated ones. In our context, a pixel can be described either by its grayvalue or by its membership to a scene object. If it is explained as part of an object, we also need to encode the presence of the object ("model cost"), as well as the error that is made by this representation. The MDL principle states that the best encoding is the one that minimizes the total description length for image, model, and error.

In accordance with the notion of description length, we can define the *savings* [14] in the encoding that can be obtained by explaining part of an image by the hypothesis h:

$$S_h = K_0 S_{area} - K_1 S_{model} - K_2 S_{error} \tag{12}$$

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. In our implementation, we assume a fixed cost $S_{model} = 1$ for each additional scene object. As an estimate for the error we use

$$S_{error} = \sum_{\mathbf{p} \in Seg(h)} (1 - p(\mathbf{p} = \textit{figure}|h)) \tag{13}$$

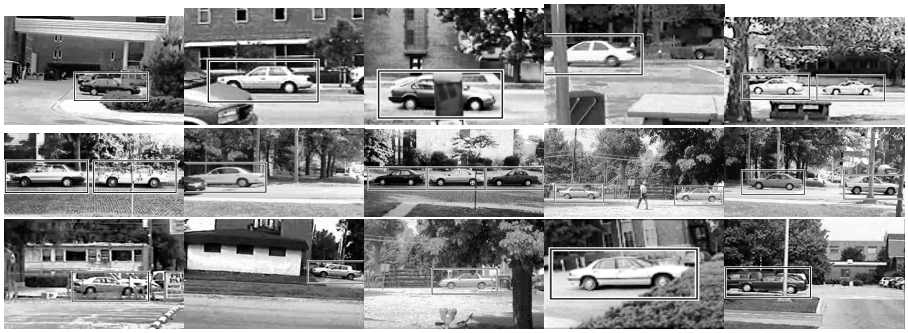


Fig. 5. *Example detections on difficult images from the test set.*

that is, over all pixels that are hypothesized to belong to the segmentation of h, we sum the probabilities that these pixels are not *figure*.

The constants K_0, K_1, and K_2 are related to the average cost of specifying the segmented object area, the model, and the error, respectively. They can be determined on a purely information-theoretical basis (in terms of bits), or they can be adjusted in order to express the preference for a particular type of description. In practice, we only need to consider the relative savings between different combinations of hypotheses. Thus, we can divide Eq(12) by K_0 and, after some simplification steps, we obtain

$$S_h = -\frac{K_1}{K_0} + (1 - \frac{K_2}{K_0})N + \frac{K_2}{K_0} \sum_{\mathbf{p} \in Seg(h)} p(\mathbf{p} = \textit{figure}|h). \tag{14}$$

This leaves us with two parameters: $\frac{K_2}{K_0}$, which encodes the relative importance that is assigned to the support of a hypothesis, as opposed to the area it explains; and $\frac{K_1}{K_0}$, which specifies the total weight a hypothesis must accumulate in order to provide any savings. Good values for these parameters can be found by considering some limiting cases, such as the minimum support a hypothesis must have in the image, before it should be accepted.

Using this framework, we can now resolve conflicts between overlapping hypotheses. Given two hypotheses h_1 and h_2, we can derive the savings of the *combined hypothesis* $(h_1 \cup h_2)$:

$$S_{h_1 \cup h_2} = S_{h_1} + S_{h_2} - S_{area}(h_1 \cap h_2) + S_{error}(h_1 \cap h_2) \tag{15}$$

Both the overlapping area and the error can be computed from the segmentations obtained in Section 3.3. Let h_1 be the stronger hypothesis of the two. Under the assumption that h_1 opaquely occludes h_2, we can set $p(\mathbf{p} = \textit{figure}|h_2) = 0$ wherever $p(\mathbf{p} = \textit{figure}|h_1) > p(\mathbf{p} = \textit{ground}|h_1)$, that is for all pixels that belong to the segmentation of h_1. Rather than to search for the globally optimal solution, which may become untractable, it is sufficient for our application to consider only pairwise combinations, as argued in [14].

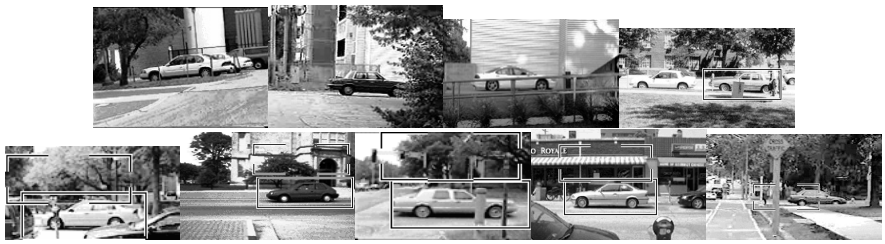


Fig. 6. *All missing detections (above) and false positives (below) our algorithm returned on the car test set. The last picture contains both a false positive and a missing detection.*

5.1 Experimental Results

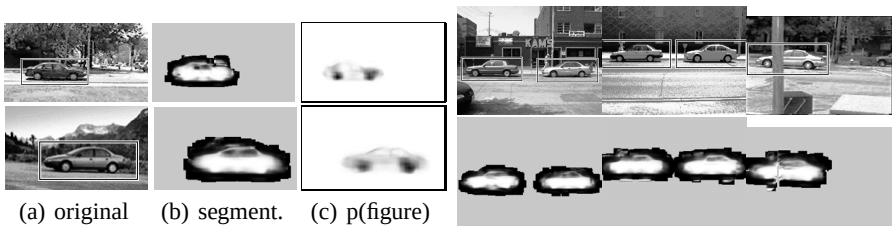
Figure 2 shows the results on the UIUC car database when the MDL criterion is applied as a verification stage. As can be seen from the figure, the results are significantly improved, and the EER performance increases from 91% to 97.5%. Without the verification stage, our algorithm could reach this recall rate only at the price of a reduced precision of only 74.1%. This means that for the same recall rate, the verification stage manages to reject 64 additional false positives while keeping all correct detections. In addition, the results become far more stable over a wider parameter range than before. This can be illustrated by the fact that even when the initial acceptance threshold is lowered to 0, the MDL criterion does not return more than 20 false positives. This property, together with the criterion's good theoretical foundation and its ability to correctly solve cases like the one in Figure 3, makes it an important contribution.

Figure 4 shows a comparison of our method's performance with other results reported in the literature. The adjacent table contains a comparison of the equal error rates (EER) with three other approaches. With an EER of 97.5%, our method presents a significant improvement over previous results. Some example detections in difficult settings can be seen in Figure 5. The images show that our method still works in the presence of occlusion, low contrast, and cluttered backgrounds. At the EER point, our method correctly finds 195 of the 200 test cases with only 5 false positives. All of these cases are displayed in Figure 6. The main reasons for missing detections are combinations of several factors, such as low contrast, occlusion, and image plane rotations, that push the object hypothesis below the acceptance threshold. The false positives are due to richly textured backgrounds on which a large number of spurious object parts are found.

In addition to the recognition results, our method automatically generates object segmentations from the test images. Figure 7 shows some example segmentations that can be achieved with this method. Even though the quality of the original images is rather low, the segmentations are reliable and can serve as a basis for later processing stages, e.g. to further improve the recognition results using global methods.

6 Recognition of Articulated Objects

Up to now, we have only considered static objects in our experiments. Even though environmental conditions can vary greatly, cars are still rather restricted in their possible



(a) original (b) segment. (c) p(figure)

Fig. 7. *(left) Example object detections, segmentations, and figure probabilities automatically generated by our method; (right) Some more detections and segmentations (white: figure, black: ground, gray: not sampled).*

shapes. This changes when we consider articulated objects, such as walking animals. In order to fully demonstrate our method's capabilities, we therefore apply it to a database of video sequences of walking cows originally used for detecting lameness in livestock [16]. Each sequence shows one or more cows walking from right to left in front of different, static backgrounds.

For training, we took out all sequences corresponding to three backgrounds and extracted 113 randomly chosen frames, for which we manually created a reference segmentation. We then tested on 14 different video sequences showing a total of 18 unseen cows in front of different backgrounds and with varying lighting conditions. Some test sequences contain severe interlacing and MPEG-compression artefacts and significant noise. Altogether, the test suite consists of a total of 2217 frames, in which 1682 instances of cows are visible by at least 50%. This provides us with a significant number of test cases to quantify both our method's ability to deal with different articulations and its robustness to occlusion. Using video sequences for testing also allows to avoid any bias caused by selecting only certain frames. However, since we are still interested in a single-frame recognition scenario, we apply our algorithm to each frame separately. That is, no temporal continuity information is used for recognition, which one would obviously add for a tracking scenario.

We applied our method to this test set using exactly the same detector settings as before to obtain equal error rate for the car experiments. The only change we made was to slightly adjust the sensibility of the interest point detector in order to compensate for the lower image contrast. Using these settings, our detector correctly finds 1535 out of the 1682 cows, corresponding to a recall of 91.2%. With only 30 false positives over all 2217 frames, the overall precision is at 98.0%. Figure 8 shows the precision and recall values as a function of the visible object area. As can be seen from this plot, the method has no difficulties in recognizing cows that are fully visible (99.1% recall at 99.5% precision). Moreover, it can cope with significant partial occlusion. When only 60% of the object is visible, recall only drops to 79.8%. Even when half the object is occluded, the recognition rate is still at 69.0%. In some rare cases, even a very small object portion of about $20 - 30\%$ is already enough for recognition (such as in the leftmost image in Figure 10). Precision constantly stays at a high level.

False positives mainly occur when only one pair of legs is fully visible and the system generates a competing hypothesis interpreting the front legs as rear legs, or vice versa. Usually, such secondary hypotheses are filtered out by the MDL stage, but if the

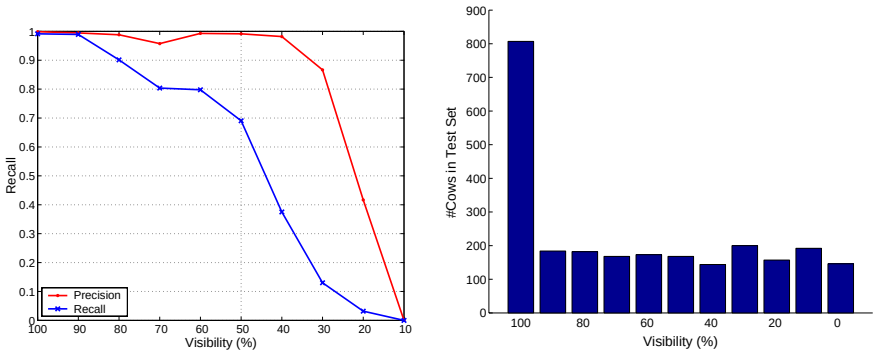


Fig. 8. *(left) Precision/Recall curves for the cow sequences when x% of the cow's length is visible. (right) Absolute number of test images for the different visibility cases.*

correct hypothesis does not have enough support in the image due to partial visibility, the secondary hypothesis sometimes wins.

Figures 9 and 10 show example detection and segmentation results for two sequences. As can be seen from these images, the system not only manages to recognize unseen-before cows with novel texture patterns, but it also provides good segmentations for them. Again, we want to emphasize that no tracking information is used to generate these results. On the contrary, the capability to generate object segmentations from single frames could make our method a valuable supplement to many current tracking algorithms, allowing to (re-)initialize them through shape cues that are orthogonal to those gained from motion estimates.

7 Discussion and Conclusion

The probabilities $p(\mathbf{p} = \textit{figure}|h)$ in Figs. 3 and 7 demonstrate why our approach is successful. These probabilities correspond to the per-pixel confidence the system has in its recognition and segmentation result. As can be seen from the figure, the cars' wheels are found as the most important single feature. However, the rest of the chassis and even the windows are represented as well. Together, they provide additional support for the hypothesis. This is possible because we do not perform any feature selection during the training stage, but store all local parts that are repeatedly encountered on the training objects. The resulting complete representation allows our approach to compensate for missing detections and partial occlusions.

Another factor to the method's success is the flexibility of representation that is made possible by the Implicit Shape Model. Using this framework, it can interpolate between local parts seen on different training objects. As a result, the method only needs a relatively small number of training examples to recognize and segment categorical objects in different articulations and with widely varying texture patterns.

The price we have to pay for this flexibility is that local parts could also be matched to potentially illegal configurations, such as a cow with 6 legs. Since each hypothesized leg is locally consistent with the common object center, there would be nothing to prevent such configurations. In our experiments, however, the MDL criterion effectively

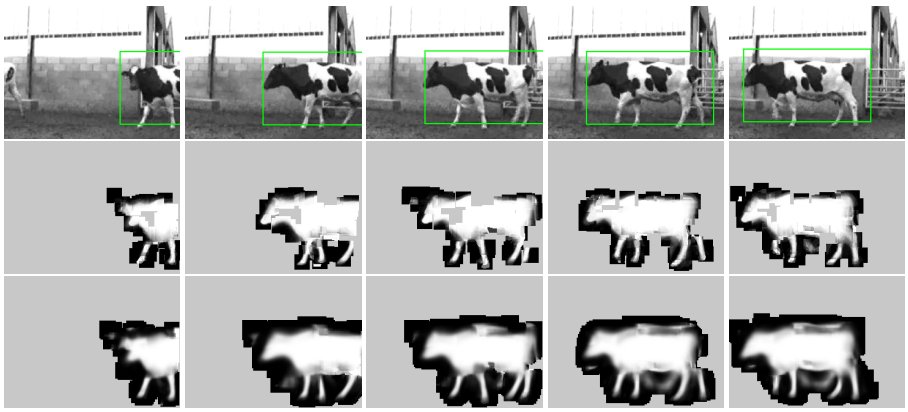


Fig. 9. *Example detections and automatically generated segmentations from one cow sequence. (middle row) segmentations obtained from the intial hypotheses; (bottom row) segmentations from refined hypotheses.*

solves this problem. Another solution would be to add a global, explicit shape model on top of our current implicit model. Using the obtained object segmentations as a guide, such a model could be learned on-line, even after the initial training stage.

Currently, our approach only tolerates small scale changes of about $10-15\%$. As our next step, we will therefore aim to extend the approach to multiple scales. Recent work by several researchers has shown considerable promise that this problem may be dealt with by using scale-invariant interest point detectors [9, 17, 8]. Also, the current model is purely representational. Although equation (5) allows for the inclusion of negative training examples, we do not yet use any such discriminative information, nor do we model the background explicitly. For the data sets used in this evaluation, this was not necessary, but we expect that the performance and robustness of our method can be further improved by incorporating these steps. Finally, we will explore how the method scales to larger object sets and how multi-view objects should best be treated.

In conclusion, we have presented a method that combines the capabilities of object categorization and segmentation in one common probabilistic framework. This paper extends our previous method by a novel hypothesis verification criterion based on the MDL principle. This criterion significantly improves the method's results and allows to handle scenes containing multiple objects. In addition, we have presented an extensive evaluation on two large data sets for cars and cows. Our results show that the method achieves excellent recognition and segmentation results, even under adverse viewing conditions and with significant occlusion. At the same time, its flexible representation allows it to generalize already from small training sets. These capabilities make it an interesting contribution with potential applications in object detection, categorization, segmentation and tracking.

Acknowledgments: This work is part of the CogVis project, funded in part by the Comission of the European Union (IST-2000-29375), and the Swiss Federal Office for Education and Science (BBW 00.0617).

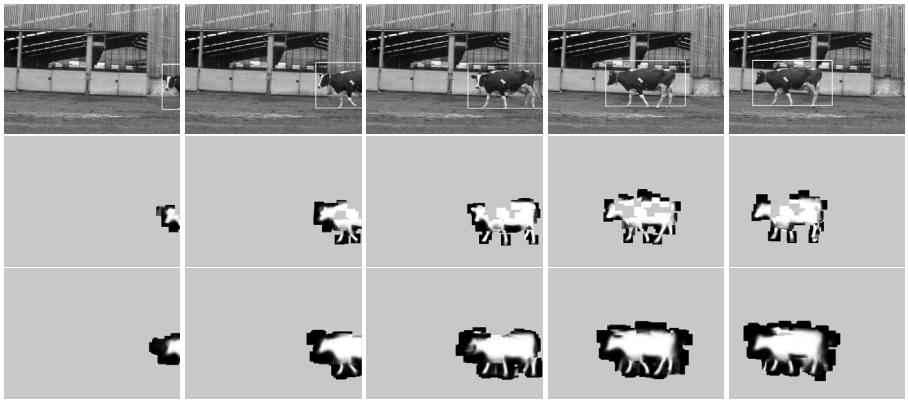


Fig. 10. *Example detections and automatically generated segmentations from another sequence. Note in particular the leftmost image, where the cow is correctly recognized and segmented despite a high degree of occlusion.*

References

1. S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV'02*, 2002.
2. D.H. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.
3. E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV'02*, LNCS 2353, pages 109–122, 2002.
4. M.C. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *ECCV'98*, 1998.
5. Y. Cheng. Mean shift mode seeking and clustering. *Trans. PAMI*, 17(8):790–799, Aug. 1995.
6. D. Comaniciu and P. Meer. Distribution free decomposition of multivariate data. *Pattern Analysis and Applications*, 2(1):22–30, 1999.
7. T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. In *ECCV'98*, 1998.
8. G. Dorko and C. Schmid. Selection of scale invariant parts for object class recognition. In *ICCV'03*, 2003.
9. R. Fergus, A. Zisserman, and P. Perona. Object class recognition by unsupervised scale-invariant learning. In *CVPR'03*, 2003.
10. A. Garg, S. Agarwal, and T. Huang. Fusion of global and local information for object detection. In *ICPR'02*, 2002.
11. C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.
12. M. Jones and T. Poggio. Model-based matching by linear combinations of prototypes. MIT AI Memo 1583, MIT, 1996.
13. B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC'03*, 2003.
14. A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation of range images as the search for geometric parametric models. *IJCV*, 14:253–277, 1995.
15. D. Lowe. Object recognition from local scale invariant features. In *ICCV'99*, 1999.
16. D. Magee and R. Boyle. Detecting lameness using 're-sampling condensation' and 'multi-stream cyclic hidden markov models'. *Image and Vision Computing*, 20(8):581–594, 2002.

17. K. Mikolajczyk, A. Zisserman, and C. Schmid. Shape recognition with edge-based features. In *BMVC'03*, pages 779–788, 2003.
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. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 38(1):15–33, 2000.
20. C. Schmid. Constructing models for content-based image retrieval. In *CVPR'01*, 2001.
21. H. Schneiderman and T. Kanade. A statistical method of 3d object detection applied to faces and cars. In *CVPR'00*, pages 746–751, 2000.
22. S. Ullman. Three-dimensional object recognition based on the combination of views. *Cognition*, 67(1):21–44, 1998.
23. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR'01*, pages 511–518, 2001.
24. M. Weber, M. Welling, and P. Perona. Unsupervised learning of object models for recognition. In *ECCV'00*, 2000.
25. S.X. Yu and J. Shi. Object-specific figure-ground segregation. In *CVPR'03*, 2003.
26. A.L. Yuille, D.S. Cohen, and P.W. Hallinan. Feature extraction from faces using deformable templates. In *CVPR'89*, 1989.