Integrating Representative and Discriminant Models for Object Category Detection

Mario Fritz[1], Bastian Leibe[1], Barbara Caputo[2], Bernt Schiele[1]

[1]Multimodal Interactive Systems, TU Darmstadt, Germany
{fritz,leibe,schiele}@informatik.tu-darmstadt.de

[2] Computer Vision and Active Perception Laboratory, KTH Stockholm, Sweden
caputo@nada.kth.se

Abstract

Category detection is a lively area of research. While categorization algorithms tend to agree in using local descriptors, they differ in the choice of the classifier, with some using generative models and others discriminative approaches. This paper presents a method for object category detection which integrates a generative model with a discriminative classifier. For each object category, we generate an appearance codebook, which becomes a common vocabulary for the generative and discriminative methods. Given a query image, the generative part of the algorithm finds a set of hypotheses and estimates their support in location and scale. Then, the discriminative part verifies each hypothesis on the same codebook activations. The new algorithm exploits the strengths of both original methods, minimizing their weaknesses. Experiments on several databases show that our new approach performs better than its building blocks taken separately. Moreover, experiments on two challenging multi-scale databases show that our new algorithm outperforms previously reported results.

1. Introduction

In recent years object categorization has regained interest and impressive results have been reported on various databases. Interestingly, while there seems to be a common consensus on the use of local features, there is much more variety on the classification methods, where the range goes from probabilistic models [7, 11] to discriminative approaches [20, 24, 15].

Generative models are quite appealing for various reasons in the context of object categorization. For example those models can be learned incrementally [19], they can deal with missing data in a principled way, they allow for modular construction of composed solutions to complex problems and therefore lend themselves to hierarchical classifier design. Also, prior knowledge can be easily taken into account [12]. In practice generative models show considerable robustness with respect to partial occlusion and viewpoint changes and can tolerate significant intra-class variation of object appearance [7, 11]. However, the price for this robustness typically is that they tend to produce a significant number of false positives. This is particularly true for object classes which share a high visual similarity such as horses and cows.

Discriminative methods enable the construction of flexible decision boundaries, resulting in classification performances often superior to those obtained by purely probabilistic or generative models [9, 14]. This allows for example to explicitly learn the discriminant features of one particular class vs. background [25] or between multiple classes [20, 15]. Object categorization algorithms which use discriminative methods combined with global and/or local representations have been shown to perform well in the presence of clutter, viewpoint changes, partial occlusion and scale variations. Also, recent work has shown the suitability of discriminative methods for recognition of large numbers of categories [20].

While so far the object recognition community has chosen one of these two modeling approaches, there has been an increasing interest in the machine learning community in developing algorithms which combine the advantages of discriminative methods with those of probabilistic generative models [9]; a similar strategy has proven to be beneficial for image parsing into regions and objects [22].

In this paper we integrate two different types of approaches into a single common framework to fully exploit their strengths while minimizing their weaknesses. More specifically, we combine the Implicit Shape Model (ISM, [11]) based on a codebook representation (which can be seen as a non-parametric probabilistic model of the appearance of object categories) with an SVM using Local Kernels

(LK, [26]), which has proven effective for object categorization [15]. The idea to use a generative model inside a kernel function has been proposed before [9, 10, 23, 21], and it has been applied to visual recognition tasks like object identification [23].

The first main contribution of this paper is a new approach which tightly integrates a probabilistic with a discriminant approach into a single categorization framework. This tight integration is made possible by a unified data representation used by both approaches. The new integrated approach is beneficial with respect to ISM, since the new approach preserves the generalization capabilities of ISM but increases its accuracy in rejecting false positives. Since ISM effectively acts as a pre-filter to the discriminant part of the algorithm the integration is also beneficial with respect to LK by using the discriminant power only where it is needed, namely on visually similar appearances of object classes.

The second main contribution are experimental results which show the superiority of the new integrated approach with respect to ISM and LK both in terms of detection performance and of a significant reduction of false positives on challenging databases. The new approach also outperforms state-of-the-art object categorization methods on challenging multi-scale data sets.

The third main contribution is that the integrated approach improves over and extends the original LK approach [26] in various respects: the new approach is scale invariant, enables localization of the object in the scene, and allows cross-instance learning of object category models.

The rest of the paper is organized as follows: after a brief review of the ISM and LK algorithms (Sec. 2), we introduce the new approach, describing in detail how it integrates ISM and LK and discussing its advantages with respect to the two previous methods (Sec. 3). Section 4 reports experiments benchmarking our new method with its building blocks, on several databases of increasing difficulty.

2. Previous Approaches

Our approach is motivated by two recent advances in object detection and discrimination.

Object Detection with Implicit Shape Models. Implicit Shape Models (ISMs) [11] are unique in that they address object category detection and top-down segmentation at the same time. They proceed by first collecting the evidence from local features in a probabilistic Hough voting procedure to determine possible object locations and scales. For each such hypothesis, they then 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. The segmentation information can then in turn be used to improve the accuracy of the detection and resolve ambiguities between overlapping hypotheses [11]. As a result of this iterative process, ISMs have been shown to yield good object detection results and considerable robustness to partial occlusion.

The ISM approach provides a flexible representation of the target category. Since each image patch votes for the object center independently of the other patches, the resulting model can interpolate between local parts seen on different training objects. As a result, it can adapt well to novel objects of the target category and typically achieves high recall. However, as a price for this flexibility, it cannot reject false positives as accurately as a discriminative model.

SVM Classification with Local Kernels. Most current object category detection systems are based on local features in order to reduce the influence of intra-class variations, noise, and occlusion [7, 2, 25, 24, 13, 11]. Support Vector Machines (SVMs), on the other hand, have shown impressive learning and recognition performance [17, 16, 8]. As the SVMs' machinery requires the computation of scalar products on the feature vectors, [26] introduced a local kernel which formulates the feature matching step as part of the kernel itself. Despite the claim in [26], this family of kernels is not a Mercer kernel [4]. Still, it can be shown that it statistically approximates a Mercer kernel in a way that makes it a suitable kernel for visual applications. On the basis of this finding, and of its reported effectiveness for object categorization [15], we will use this family of kernels in this paper.

Given two local feature lists L_h and L_k, these local kernels are defined as [26]

$$K(L_h, L_k) = \frac{1}{n_h} \sum_{j_h=1}^{n_h} \max_{j_k=1,\ldots,n_k} \left\{ K_l(L_h^{j_h}, L_k^{j_k}) \right\}, \quad (1)$$

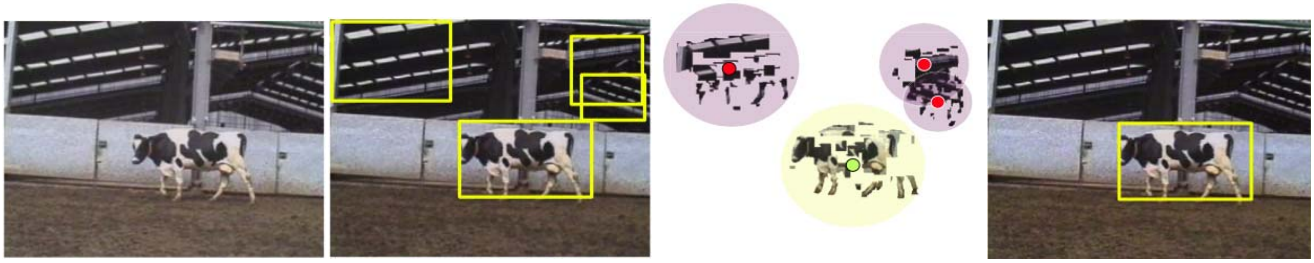
where the local feature similarity kernel K_l consists of an appearance part K_a and a position constraint K_p

$$K_l(L_h^a, L_k^b) = K_a(L_h^a, L_k^b)\ K_p(pos(L_h^a), pos(L_k^b)). \quad (2)$$

Various options have been given for the selection of K_a and K_p [26], including the following choice

$$K_a = exp\left\{ -\gamma \left(1 - \frac{\langle x - \mu_x | y - \mu_y \rangle}{||x - \mu_x|| ||y - \mu_y||} \right) \right\} \quad (3)$$

As shown by [26, 15], Local SVMs can discriminate well between different object categories. However, they contain no localization component and require accurate initialization in position and scale. In the literature, the standard solution to this problem is to perform an exhaustive search over all possible object positions and scales [16, 18, 8, 24, 2, 20]. However, this exhaustive search imposes severe constraints, both on the detector's computational complexity and on its discriminance, since a large number of potential false positives need to be excluded. In this paper, we present a different solution to this problem by integrating Local SVMs with the ISM approach.



(a) original image (b) detected hypotheses (c) SVM input (d) accepted hypotheses

Figure 1. *Stages of the integrated approach. (a) original image; (b) hypotheses detected by the representative ISM; (c) input to the SVM stage; (d) verified hypotheses.*

3. Integrated Approach

The main contribution of this paper is to integrate both approaches into a consistent framework (visualized in Fig. 1). Applied to a novel test image (Fig. 1(a)), the representative ISM is first used to find a set of promising hypotheses (Fig. 1(b)) and estimate their support in both location and scale (Fig. 1(c)). For each of those hypotheses, the more exact discriminative model is then applied in order to verify them and filter out false positives (Fig. 1(d)). By using the same internal representation, namely the appearance codebooks, those two approaches are tightly integrated. The ISM uses these appearance codebooks to generate hypotheses which are visually consistent and which follow a weak spatial model. The discriminative model on the other hand uses the same appearance codebooks to find visually discriminant information for object classes and also to add a stronger spatial model effectively extracting discriminant spatial codebook distributions. We thus combine the capabilities of both models in an advantageous manner.

3.1. Generation of an Appearance Codebook

As a common representation, we generate a category-specific appearance codebook, as described in [11]. For this, we apply a scale-invariant DoG interest point operator [13] to all training images and extract image patches with a radius of 3σ of the detected scale. All extracted patches are then rescaled to a uniform size (in our case 25×25 pixels) and grouped using an agglomerative clustering scheme. The resulting clusters form a compact representation of local object structure. In the following, we keep only the cluster centers $C = (\vec{c}_1, \ldots, \vec{c}_R)$ as codebook entries.

For each codebook entry, we then learn its spatial occurrence distribution 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 $\vec{c}_i$ all locations where it can be matched to the extracted patches (where patch similarity is measured by normalized correlation).

3.2. Initial Hypothesis Generation

In order to generate initial hypotheses about possible object locations and scales, we use a scale-invariant version of the ISM approach from [11]. The approach starts by applying the same patch extraction procedure as before, 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.

This is formalized as follows. Let $\vec{e}$ be an image patch observed at location ℓ. Each matching codebook entry $\vec{c}_i$ generates probabilistic votes for different object categories o_n and locations $\lambda = (\lambda_x, \lambda_y, \lambda_s)$ according to the following marginalization:

$$P(o_n, \lambda|\vec{e}, \ell) = \sum_i P(o_n, \lambda|\vec{c}_i, \ell) p(\vec{c}_i|\vec{e}) \quad (4)$$

where $p(\vec{c}_i|\vec{e})$ denotes the probability that $\vec{e}$ matches to $\vec{c}_i$, and $P(o_n, \lambda|\vec{c}_i, \ell)$ describes the stored spatial probability distribution for the object center relative to an occurrence of that codebook entry. For describing the matching probability, we make the assumption that an image patch can be approximated by the mean of the closest-matching codebook entries $C^*_{\vec{e}} = \{\vec{c}^*_i | sim(\vec{c}^*_i, \vec{e}) \geq \theta\}$, thus setting $p(\vec{c}^*_i|\vec{e}) = \frac{1}{|C^*_{\vec{e}}|}$. Object hypotheses are found as maxima in the 3D voting space using Mean-Shift Mode Estimation [5] with a scale-adaptive *balloon density estimator* [6] and a uniform ellipsoidal kernel K:

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

Once a hypothesis has been found, the contributing votes are backprojected to determine which local features and codebook activations supported it in the image (Fig. 1(c)). The original ISM approach additionally computes a full top-down segmentation of the object, which has been shown to improve the results considerably. In our approach, we also

apply this segmentation loop to improve the quality of hypotheses. However, as this part is of minor relevance to the understanding of our integrated approach, we refer the reader to [11] for details.

3.3. Representation in Codebook Coordinates

The result of the ISM stage is a set of object hypotheses $h = (o_n, \lambda)$, together with their support in the image (Fig. 1(c)). This support consists of a list of local features that contributed to the hypothesis and their corresponding codebook activations. In order to interpret this information in the SVM framework, we first have to adapt the kernel formulation to our codebook representation.

The key idea is that the scalar product $\langle \vec{x}, \vec{y} \rangle$ used in the SVM Kernel can be expressed in terms of a codebook matching problem. For this, we project both $\vec{x}$ and $\vec{y}$ into the affine space spanned by the codebook entries $\vec{c}_i$ as basis vectors. With $\vec{x} = \sum_i a_i \vec{c}_i$ and $\vec{y} = \sum_j b_j \vec{c}_j$ the scalar product can be written as

$$\langle \vec{x}, \vec{y} \rangle = \sum_i \sum_j a_i \langle \vec{c}_i, \vec{c}_j \rangle b_j. \quad (5)$$

This formulation has two advantages. One is its computational efficiency – both the intra-codebook similarity matrix $\langle \vec{c}_i, \vec{c}_j \rangle$ and the support vector coefficients b_j can be precomputed. Only the image-feature coefficients a_i need to be calculated during recognition. The second advantage is that the data is now expressed in a common format and partial results can be reused by both stages.

Remains the problem how to select the coefficients a_i and b_j. The smallest reconstruction error would be obtained by a least-squares solution, but this solution is typically not sparse. In order to arrive at a sparse representation, we again consider only the closest-matching codebook entries $C^*_{\vec{x}} = \{\vec{c}^*_i | sim(\vec{c}^*_i, \vec{x}) \geq \theta\}$ and approximate the vectors $\vec{x}$ and $\vec{y}$ by the mean of those "activated" codebooks. Thus, with $n = |C^*_{\vec{x}}|$, $m = |C^*_{\vec{y}}|$, we arrive at

$$\langle \vec{x}, \vec{y} \rangle \approx \langle \vec{\mu}_x, \vec{\mu}_y \rangle = < \frac{1}{n} \sum_{i=1}^{n} \vec{c}^*_i, \frac{1}{m} \sum_{j=1}^{m} \vec{c}^*_j > \quad (6)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{n} \langle \vec{c}^*_i, \vec{c}^*_j \rangle \frac{1}{m}. \quad (7)$$

This approximation is justified under the assumption that the codebook entries sufficiently cover the relevant "object" region of the appearance space. We have verified the validity of this assumption in a series of control experiments. The results indicate that the difference in reconstruction error between the least-squares solution and our sparse approximation is only modest and subsumes to an average error of approximately one gray level per pixel on a reconstructed patch. As a result, we get a problem-specific representation which expresses the data in a common vocabulary and is used throughout both stages of our approach. In particular, this representation allows us to reuse the results of the initial codebook matching stage for the SVM model.

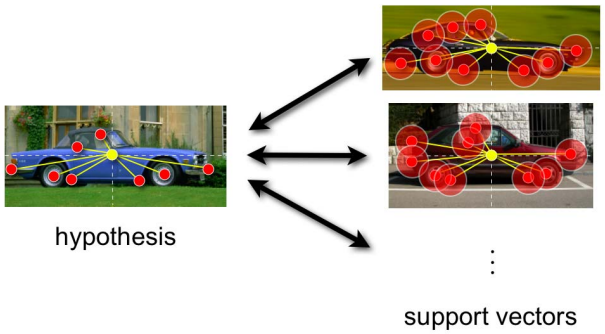


Figure 2. *A look inside the LK verification stage. Each support vector specifies a configuration of local features, corresponding to a particular training example. When evaluating a hypothesis, the kernel K first searches for the k best feature correspondences, considering both appearance and relative position, and uses them to judge the quality of the match.*

3.4. SVM Verification with Local Kernels

Let $X = \{(x_1, \lambda_1), \ldots, (x_N, \lambda_N)\}$ be a set of local features (with appearance and relative location) supporting hypothesis h, and $A = \{A_1, \ldots, A_N\}, A_i = (a_1, \ldots, a_R)$ be their corresponding codebook activations. The ISM procedure guarantees that each feature in the supporting set is consistent in appearance and location with at least one training example. However, as only local consistency is enforced, this reference example may be a different one for each feature. In the next step, we therefore want to verify that the global feature configuration is also consistent.

Figure 2 visualizes the chosen verification procedure. In the remainder of this section, we will define the Local Kernel in a way that each support vector corresponds to a distinct configuration of local features. When evaluating a hypothesis, it is successively compared to each support vector. For each such match, correspondences are established between visually similar features occurring in the same relative locations (with a small tolerance σ), and the quality of the resulting global configuration fit is measured. Thus, the kernel enforces strong spatial constraints to verify the hypothesis.

This is done as follows. Let $Y = \{(y_1, \lambda_1), \ldots, (y_M, \lambda_M)\}$ be the features observed on a training image with corresponding codebook activations $B = \{B_1, \ldots, B_M\}, B_j = (b_1, \ldots, b_R)$. In order to compare the feature configurations of X and Y, we first try to find a set of correspondences between their features. For each pair of features $(\vec{x}, \lambda_x)$ and $(\vec{y}, \lambda_y)$, the quality of a

match is expressed by the local similarity kernel K_l:

$$K_l((\vec{x},\lambda_x),(\vec{y},\lambda_y)) = K_a(\vec{x},\vec{y})\; K_p(\lambda_x,\lambda_y), \tag{8}$$

where K_a is measuring the appearance similarity and K_p is imposing a position constraint in the manner of a penalty function. For K_a and K_p we stick to the choices made in [26], but replace the correlation coefficient by the approximation from eq. (5):

$$\begin{aligned} K_a(\vec{x},\vec{y}) &= \exp\left(-\gamma\left(1-\langle\vec{x},\vec{y}\rangle\right)\right) \\ &\approx \exp(-\gamma(1-\sum_i\sum_j a_i\,\langle\vec{c}_i,\vec{c}_j\rangle\,b_j)) \end{aligned} \tag{9}$$

$$K_p(\lambda_x,\lambda_y) = \exp\left(-\frac{(\lambda_x-\lambda_y)^2}{2\sigma^2}\right). \tag{10}$$

In order to allow for some flexibility in the part arrangement, we do not enforce complete correspondence, but only match a subset of the features by searching for the k best correspondences. This is done using a greedy selection strategy on the feature similarity matrix $K_l(X,Y)$. Let $\Phi \in \pi_1^N$, $\Psi \in \pi_1^M$ be permutations of the local features to reflect this greedy assignment. According to [26], the corresponding Local Kernel is then defined as

$$K(X,Y) = \frac{1}{k}\max_{\Phi,\Psi}\sum_{j=1}^{k} K_l\left((\vec{x}_{\Phi(j)},\lambda_{x,\Phi(j)}),(\vec{y}_{\Psi(j)},\lambda_{y,\Psi(j)})\right). \tag{11}$$

Note that the resulting kernel does not need to consider the original features anymore, but only operates on the codebook activations passed from the previous stage. It thus requires very little computation and imposes only a small overhead on the total execution time. In all experiments presented in this paper, we set k to 50 and determine the remaining parameters using cross-validation on the training set.

3.5. Discussion

It is important to emphasize that through the integration, the SVM stage is solving a simpler problem than the previous LK approach. Not only is it initialized with an estimate of the object position and scale, but it directly obtains also the supporting image features as input. It can thus optimize its decision surface on the failure cases of the ISM stage and learn a stronger discriminative model. In addition, the discriminative model makes it possible to achieve a better separation from background constellations, whose complex distributions are notoriously hard to express in a probabilistic framework.

The matching to a common codebook enables both stages to make use of "across-instance" learning which is essential when dealing with limited training set sizes. In addition, the Local Kernel stage complements the ISM's weak spatial model with stronger spatial constraints.

As a side benefit, the output confidence of the SVM stage (i.e. the distance to the hyperplane) becomes comparable for different object categories. This is the case because the Local Kernel bases its computation on a fixed number of k correspondences.

4. Experiments

In this section, we show that our new approach benefits from the integrated representative and discriminative representation (in the following termed IRD). We present our results in three steps. First, we compare our new approach to the original ISM and LK approaches. Section 4.2 then reports results on a multi-category detection/discrimination task. Finally, we evaluate our approach on two difficult data sets containing large scale changes and partial occlusion.

Data. In order to evaluate our approach, we apply it to a test set containing objects of four categories, namely cars, cows, horses, and motorbikes. The pairs cars/motobikes and cows/horses were especially chosen to measure cross-category confusions, since they share similar visual features. The data is mostly taken from the PASCAL database collection [1]. For cars we use the UIUC single-scale test set; for motorbikes the CalTech motorbike set (with the same training/test split as in [7]); and for cows the TUD cow database (supplemented with 556 test images) . For the background set, we use 450 CalTech background images. The horse images are taken from the Weizmann horse database [3] and split into 79 training and 164 test images. This is the first time detection results are reported on this database, as it was previously only used for segmentation tasks.

4.1. Comparison with Original Approaches

We start the experimental part with a comparison of our new IRD approach with the approaches it originates from – namely the Local Kernels and the ISM. To provide a fair comparison with the LK approach, which is not designed to be scale invariant or perform a detection task in the first place, we report results of object present/absent experiments. The test is performed on images of each category vs. 450 novel background images.

Tabel 1 summarizes the equal error rate (EER) performances for this experiment. As can be seen from the table, the integrated approach achieves superior performance compared to its building blocks.

4.2. Multi-Category Discrimination

Detection Task and Evaluation. Our main experiments are performed on a detection task, where the detector has

	LK	ISM	IRD
car	61.0 %	94.7 %	**99.4**%
cow	95.3 %	96.1 %	**97.1**%
horse	77.8 %	88.5 %	**88.5**%
motorbike	87.6 %	93.8 %	**96.5**%

Table 1. *Equal error rate performances achieved by the Local Kernel (LK), Implicit Shape Model (ISM), and our integrated approach (IRD) on present/absent tasks.*

to localize image regions in which an instance of the category of interest is present. For evaluating the car detections, we use exactly the same acceptance criterion and evaluation software as in [2]. However, as this criterion is only well-defined for fixed-size bounding boxes (and thus not directly applicable to the cow and horse categories), we apply an extended criterion for the other three categories. We inscribe an ellipse in the ground truth bounding box and measure the distance d_r between the bounding box centers relative to the ellipse's radius at the corresponding angle. A hypothesis is accepted if $d_r \leq 0.5$ and the ground truth and hypothesis bounding boxes cover one another by at least 50%. In accordance to [2], only one hypothesis per object is accepted as correct – any additional hypothesis on the same object is counted as false positive.

Detection Results. Figure 3 shows the results of this evaluation in the form of Recall-Precision curves (RPCs). To vary the strictness of the local kernel SVM in our new IRD approach without retraining, we used the distance to the decision boundary as a confidence measure. Although the ISM by itself performs already quite well on all four categories, our new IRD approach improves the EER performance for cars from 87.6% to 88.6%, for cows from 92.5% to 93.2% and for motorbikes from 80.0% to 84.0%. For horses, the performance stays at the same level. Besides the gain in EER performance, cars, cows and motorbikes profit from the added discriminance in terms of increased precision of the final detector, which shifts the precision-recall curves to the left. Especially the relatively rigid car and motorbike categories profit from the stronger structured constraint of the local kernel. Figure 4 displays some example detections of our approach which illustrate the generalization capabilities over large intra-category variations, including different articulations, and its robustness to partial occlusion.

Discrimination Results. Given these detectors operating at their equal error rate, we now investigate the produced confusions. Table 2 displays the false detections each of the detectors produces per image on all four object categories. The left number reports the false positives detected by the ISM. It can be seen that the ISM performs well for the car model, but still produces a relatively large number of false positives for the other categories. The larger number of confusions on the car images can be explained by the fact that those images are about twice as large as the other images. The right number reports the false positives detected by our new IRD approach processing all detectors in parallel and acting as a single unified detector. Ambiguous detections are eliminated using the local SVM output as a confidence measure. We can observe a drastic reduction of false positives down to (or even below) the 0.1 level for almost all combinations. In particular, these results show that in our new IRD approach the SVM output is well suited as a confidence measure for comparing hypotheses across categories.

4.3. Discriminant Category Detection

In this section, we evaluate our approach on two more challenging databases that include large scale changes and significant partial occlusion. We use the UIUC multi-scale cars and the TUD motorbikes, which are also both part of the PASCAL collection [1]. For the multi-scale cars, we again use the acceptance criterion from [2]; for the motorbikes we use the criterion described in Section 4.2 for the reasons given there. Figure 5 shows the result of this evaluation. The black line corresponds to the performance reported by [2], with an EER performance of about 45%. In contrast, our IRD approach achieves 87.8% EER – an improvement of over 40%! Interestingly, our method obtains up to 64% recall before generating any false positives. On the motorbike test set, our approach achieves an EER performance of 81%. Compared to ISM, there is a consistent improvement in precision. The difficulty of the task is illustrated by Figure 6, which shows example detections of our IRD approach documenting the performance under occlusion, extreme illumination conditions and large scale changes.

5. Summary and Conclusions

Summarizing our approach, we integrated a representative object detection method with a discriminative verification stage for the generated hypotheses. Both stages operate on a common codebook representation. They share and reuse the same information from sampled image patches, but interpret it in different ways. The ISM hypothesis generation stage searches for agglomerations of image patches that are locally consistent with a common object center. Treating each sampled patch independently, it can interpolate between different training examples and adapt to novel objects and changed articulations. The SVM verification stage, on the other hand, enforces stronger spatial constraints and verifies the global feature configuration. At the same time, its discriminative capabilities obviate the need for a dedicated background model, which is difficult to estimate reliably in a probabilistical framework. Finally, the tight integration with the output of the

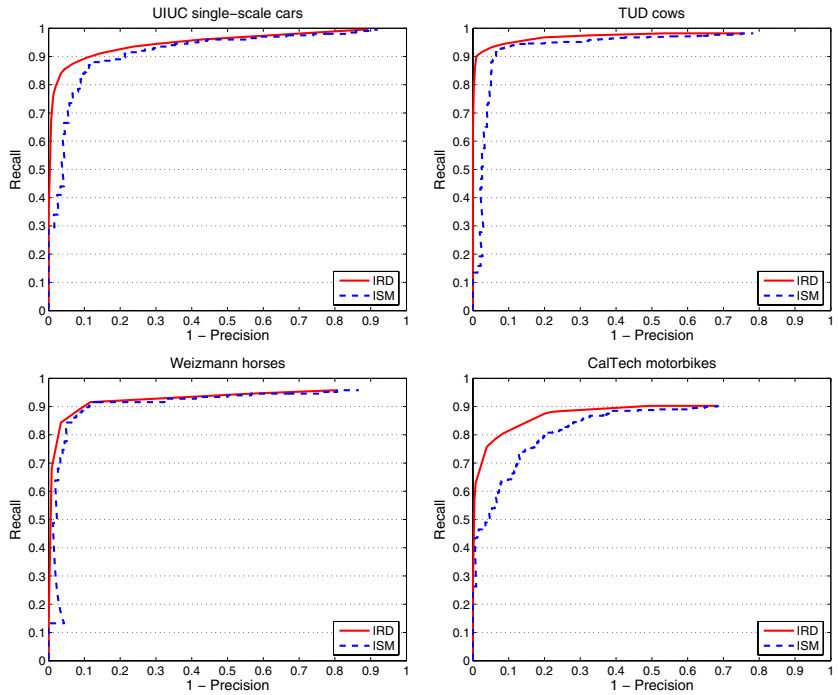


Figure 3: *Recall-Precision curves for the car, cow, horse and motorbike model on a detection task.*

ISM **IRD**	car #170		cow #557		horse #164		motorbike #400	
car	-		0.07	**0.00**	0.02	**0.01**	0.18	**0.03**
cow	1.00	**0.49**	-		0.18	**0.11**	1.05	**0.05**
horse	0.71	**0.16**	0.53	**0.08**	-		0.68	**0.05**
motorbike	1.07	**0.08**	0.29	**0.09**	0.22	**0.00**	-	

Table 2: *Cross-category confusions (false positives per test image) for the ISM and our new IRD approach on a detection task.*

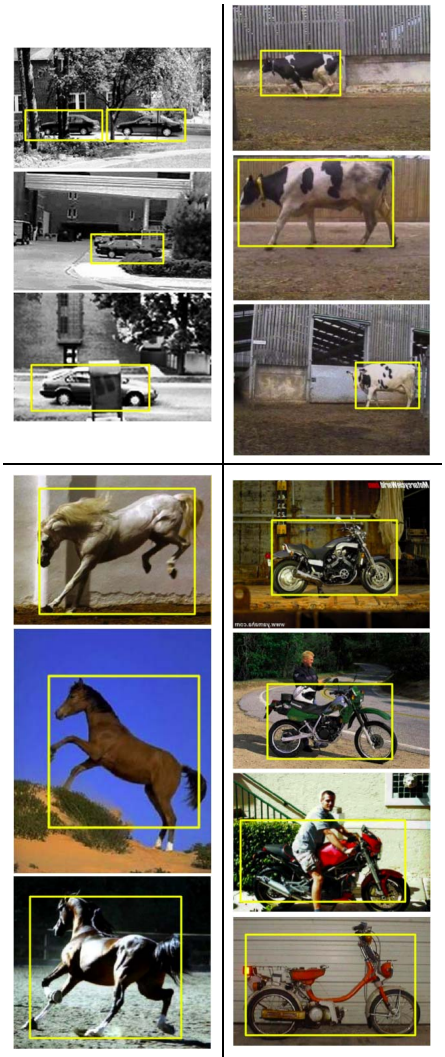


Figure 4: *Example detections on the car, cow, horse, and motorbike test sets.*

ISM stage removes the influences of translation and scale changes, which greatly simplifies the discrimination problem.

Together, both stages manage to reduce the number of false positives and cross-category confusions significantly and perform considerably better than either stage alone. In our experiments, we have shown this improvement both for a four-class detection/discrimination task and for object detection on two challenging data sets containing large scale changes and partial occlusion. In particular these last results show a considerable improvement on the state-of-the--art with over 40% difference in EER performance for the UIUC multi-scale car database.

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

References

[1] The PASCAL Object Recognition Database Collection. http://www.pascal-network.org/challenges/VOC.

[2] S. Agarwal, A. Atwan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *PAMI*, 2004.

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

[4] S. Boughorbel, J.-P. Tarel, and F. Fleuret. Non-mercer kernels for svm object recognition. In *BMVC'04*, pages 137 – 146, London, England, 2004.

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

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

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

[8] B. Heisele, T. Serre, M. Pontil, and T. Poggio. Component-based

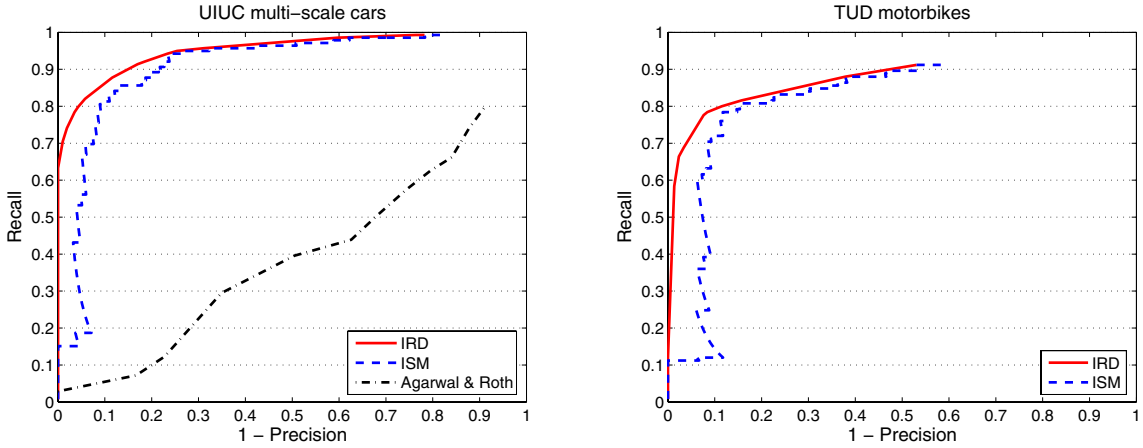


Figure 5. *Precision-Recall curves for the difficult multi-scale databases of cars and motorbikes. Our new IRD approach clearly outperforms the state-of-the-art on the UIUC multi-scale car database.*

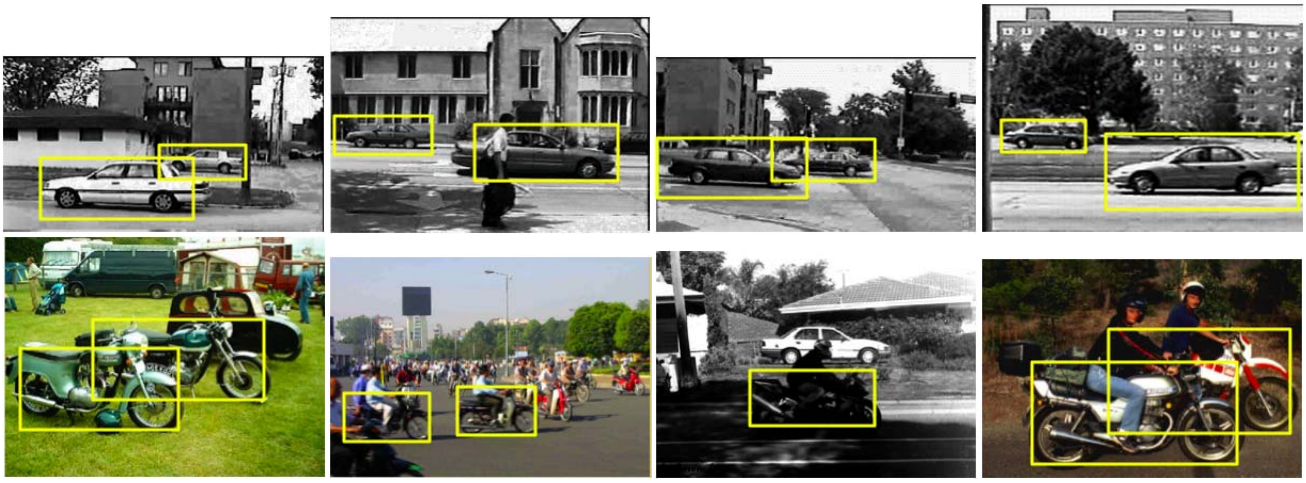


Figure 6. *Example detections of our new IRD approach on the difficult multi-scale UIUC car database and the multi-scale TUD motorbike test set.*

face detection. In *CVPR'01*, pages 657–662, 2001.

[9] T. S. Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *NIPS'98*, pages 487–493, 1998.

[10] T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *JMLR'04*, pages 819–844, 2004.

[11] 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.

[12] F. Li, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *ICCV'03*, 2003.

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

[14] A. Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *NIPS'01*, pages 841–848, 2001.

[15] M.E. Nilsback and B. Caputo. Cue integration through discriminative accumulation. In *CVPR'04*, 2004.

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

[17] M. Pontil and A. Verri. Support vector machines for 3d object recognition. *PAMI*, 20(6):637–646, 1998.

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

[19] D. Skočaj and A. Leonardis. Weighted and robust incremental method for subspace learning. In *ICCV'03*, pages 1494–1501, 2003.

[20] A. Torralba, K. Murphy, and W. Freeman. Sharing features: Efficient boosting procedures for multiclass object detection. In *CVPR'04*, 2004.

[21] K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg, and K.-R. Müller. A new discriminative kernel from probabilistic models. *Neural Computation*, 14(10):2397–2414, 2002.

[22] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. In *ICCV'03*, 2003.

[23] N. Vasconcelos, P. Ho, and P. J. Moreno. The kullback-leibler kernel as a framework for discriminant and localized representations for visual recognition. In *ECCV'04*, pages 430–441, 2004.

[24] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

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

[26] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *ICCV'03*, 2003.