Generic Object Recognition with Boosting

Andreas Opelt, Axel Pinz, *Member*, *IEEE*, Michael Fussenegger, and Peter Auer

Abstract—This paper explores the power and the limitations of weakly supervised categorization. We present a complete framework that starts with the extraction of various local regions of either discontinuity or homogeneity. A variety of local descriptors can be applied to form a set of feature vectors for each local region. Boosting is used to learn a subset of such feature vectors (weak hypotheses) and to combine them into one final hypothesis for each visual category. This combination of individual extractors and descriptors leads to recognition rates that are superior to other approaches which use only one specific extractor/descriptor setting. To explore the limitation of our system, we had to set up new, highly complex image databases that show the objects of interest at varying scales and poses, in cluttered background, and under considerable occlusion. We obtain classification results up to 81 percent ROC-equal error rate on the most complex of our databases. Our approach outperforms all comparable solutions on common databases.

Index Terms—Boosting, object categorization, object localization.

1 Introduction

Object recognition has been a long standing goal of computer vision research. Many significant contributions discuss the recognition of *specific*, individual objects from images. *Generic* object recognition is the task of classifying an individual object to belong to a certain category, thus also termed *object categorization*. While humans are quite good in categorization tasks—they often perform even better than in the recognition of individuals, the opposite is true for today's artificial vision systems. Only very recently, first success has been reported in object categorization. This success is strongly related to new algorithms which efficiently describe local, salient regions in images. At the same time, a number of common databases have been established for the sake of comparison of the emerging categorization algorithms.

There are many possible approaches to generic object recognition: Learning of constellations of local features from still images [8], [41] integration of motion cues and local features [38], and more complex geometric models (e.g., pictural structures [7]), to mention just a few. Another important issue is the amount of supervision which is supplied during the training of a recognition system. To give an example, Agarwal and Roth [1] present small rectangular regions which contain just the object of interest (a car viewed from the side). The selection of training and test images is a further, more implicit source of supervision. Most existing databases for object categorization show the objects at prominent scales, often with little background clutter, occlusion, or variance in object pose.

This paper sets out to explore the limits of weakly supervised object categorization from still images. To keep this effort tractable, we have to assume a number of boundary conditions. Most important, we focus on local descriptors of regions of discontinuity or homogeneity, without taking any spatial relations into account. Furthermore, we assume that the performance of individual descriptors might be category-specific. Thus, we use Boosting as a learning technique that can elegantly form category-specific vectors of very diverse descriptors. Finally, we define the tolerable amount of supervision by labeling the training images of the database. The system knows whether a training image contains an instance of a certain category, or not. But, it has to learn all other relevant information without further supervision (object pose, scale, and localization) and to deal with potential occlusion, varying illumination and background clutter.

The paper sheds light on the following questions: Is the performance of individual descriptors category specific? To what extent do combinations of diverse descriptors improve the categorization performance? What is learned by the system, in terms of category description *and* object localization? We also compare our approach with others based on the use of common databases. The major contributions of the paper are: 1) We present a complete framework for weakly supervised categorization. 2) We have designed publicly available new and complex databases. 3) We give experimental results for the combination of diverse local descriptors and their localization with respect to object/background in the image. Further contributions include a new similarity-measure-based segmentation algorithm and the specific application of Boosting as a popular learning technique.

The paper is organized as follows: We discuss related work in Section 2. Section 3 gives a detailed overview of our approach and explains the differences between our new data set and existing databases for object categorization. In Section 4, we present the various methods of region detection used in our framework focusing on the new Similarity-Measure-Segmentation. The local descriptors of these regions are described in Section 5. Section 6 presents our general learning approach and the combination of various kinds of description vectors. Section 7 describes our experimental setup, presents experimental results, and compares them with other approaches for object recognition. Section 8 concludes with a discussion and an outlook on further extensions.

- A. Opelt, A. Pinz, and M. Fussenegger are with the Institute of Electrical Measurement and Measurement Signal Processing, Graz University of Technology, Schiesstattg. 14b, A-8010 Graz, Austria. E-mail: {opelt, pinz, fussenegger}@tugraz.at.
- P. Auer is with the Institute for Computer Science, University of Leoben, Franz-Josef-Straße 18, A-8700 Leoben, Austria. E-mail: auer@unileoben.ac.at.

Manuscript received 12 July 2004; revised 18 Apr. 2005; accepted 20 June 2005; published online 13 Jan. 2006.
Recommended for acceptance by R. Basri.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0346-0704.

0162-8828/06/$20.00 © 2006 IEEE Published by the IEEE Computer Society

2 Related Work

Taking a closer look at the extensive body of literature on object recognition, each approach has its specific merits and limitations. In general, common approaches use image databases which show the object of interest at prominent scales and with only little variation in pose (e.g., [8], [1], [20]). Others presegment the object manually (e.g., [6], [37]) to reduce complexity. Subsequently, we discuss some of the most relevant and most recent results related to our approach and point out the differences to our method. One main extension of our approach to the existing solutions is that we do not use just one technique of information extraction, but a combination of various methods.

Boosting was successfully used by Viola and Jones [38] as the learning ingredient for a fast face detector. The weak hypotheses were the thresholded average brightnesses of collections of up to four rectangular regions. Recently, Viola et al. [39] extended this approach by also incorporating motion information. We also use different sources of information in one system, but instead of motion, we combine various region description methods in one classifier. Furthermore, Viola's work requires manually presegmented objects in their training sequences, whereas our training images are highly complex and no object segmentation is given. Schneiderman and Kanade [33] use Boosting to improve an already complex classifier. Contrary to them, we are using Boosting to combine rather simple classifiers by selecting the most discriminative features. Additionally, Schneiderman and Kanade undertake rather specific object recognition as they train each object from different viewpoints.

Also, a wide variety of other learning techniques has been used to solve the task of object recognition. For example, Agarwal and Roth [1] use Winnow as the underlying learning algorithm for the recognition of cars from side views. For this purpose, images are represented as binary feature vectors. These feature vectors encode which image patches from a "codebook" appear in an image. The bits of such a feature vector can be seen as the result of weak classifiers, one weak classifier for each position in the binary vector. For learning, it is required that the output of all weak classifiers is calculated a priori. In contrast, Boosting only needs to find the few weak classifiers which actually appear in the final classifier. This substantially speeds up learning, if the space of weak classifiers carries a structure that allows the efficient search for discriminative weak classifiers. A simple example is a weak classifier which compares a real valued feature against a threshold. For Winnow, one weak classifier needs to be calculated for each possible threshold a priori,[1] whereas for Boosting the optimal threshold can be determined efficiently when needed. The idea of Agarwal and Roth was picked up by Leibe et al. [20], who use this codebook of appearance and add an implicit shape model. This gives good classification results as well as the segmentation of the object. But, in their approach, as in the work of Agarwal and Roth, the authors manually crop out the objects for training to reduce complexity.

Wallraven et al. [40] use support vector machines combined with local features for object recognition. But, they perform a rather specific recognition task on images of lower complexity without any background clutter.

A different approach to object class recognition is presented by Fergus et al. [8]. The authors use the constellation model first proposed by Leung et al. [21] and the EM-type learning framework of Weber et al. [41] to learn this probabilistic model, but they add scale invariance to the framework. In [9], the same authors extend the constellation model to include heterogeneous parts consisting of curve segments and appearance patches. The parts and their constellations can be learned without supervision and from cluttered images. In contrast, we use a model-free approach and propose Boosting as a very different learning algorithm compared to EM.

Recently, LeCun et al. [19] studied the use of various popular learning techniques for the categorization of images with complex variabilities (clutter, varying pose, and lighting). They pointed out the limits of nearest neighbor methods and support vector machines on difficult data. Additionally, they presented promising results on a complex data set using convolutional networks. In contrast to their work, we use Boosting as learning technique. We also use local description methods instead of their global image representation via PCA.

Another object recognition approach was introduced by Dorko and Schmid [6]. It is based on the construction and selection of scale-invariant object parts. These parts are subsequently used to learn a classifier. The authors show a robust detection under scale changes and variations in viewing conditions, but in contrast to our approach, the objects of interest are manually presegmented. This dramatically reduces the complexity of distinguishing between relevant patches on the objects and background clutter.

Ferrari et al. [10] present an approach where object recognition works even if aggravating factors like background clutter, scale variations, or occlusion are very strong. Based on a model of a specific object, an iterative approach is applied. Starting with a small initial set of corresponding features good results are obtained. While this work presents a powerful concept of an iterative "active exploration" approach, it is based on a model for a specific object which is learned from noncluttered representations of the object. Another interesting approach was introduced by Selinger and Nelson [34] who perform object recognition in cluttered backgrounds. But, their approach also deals with specific objects rather than generic object categories.

A new possibility of describing objects for categorization is introduced by Thureson and Carlsson in [37]. It is based on histograms of qualitative shape indices. These indices are calculated from the combinations of triplets of location and gradient directions of the samples. The object categories are represented by a set of the histogram representations of the training images. For each new test image, the inner products of the representation vector (histogram) with all trained histograms are calculated. The smallest of these products and a threshold are used to categorize this certain image. This approach is based on a matching of image representations, whereas we compute a classifier from all the training images. This solution also requires a manual presegmentation of the relevant object to reduce complexity.

Carbonetto et al. [3] present an approach for contextual object recognition based on a segmented image. They attach labels to image regions and learn a model of spatial relationships between them. We also use segments as image representations, but we can cope with more complex images using our model-free approach.

[1] More efficient techniques for Winnow like using virtual threshold gates [24] do not improve the situation much.

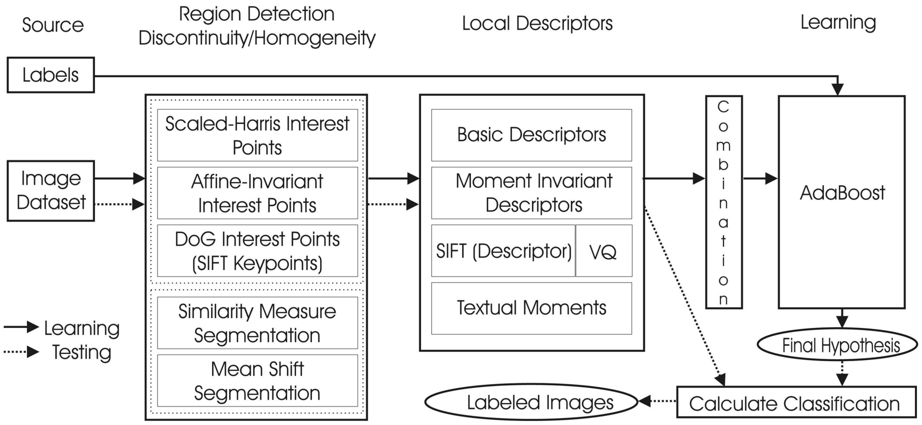


Fig. 1. Our framework for generic object recognition starts from a labeled image database. Regions of discontinuity and homogeneity are extracted and described by local descriptors forming a feature vector. Learning by AdaBoost [12] leads to a final hypothesis which consists of several weak hypotheses. The solid arrows show the training procedure, the dotted ones the testing procedure.

3 Method and Data

To learn a category, the learning algorithm is provided with a set of labeled training images. A positive label indicates that a relevant object appears in the image. The objects are not presegmented, their location in the images and the viewpoints are unknown. As output, the learning algorithm delivers a final classifier (further on also called "final hypothesis") which predicts if a relevant object is present in a new image. The learning procedure in our framework (see Fig. 1) works as follows: The labeled images are put through a preprocessing step that transforms them to gray scale.[2] Then, two kinds of regions are detected. On the one hand, regions of discontinuity are extracted. These are regions around salient points normalized to quadratic patches. They are extracted with various existing methods. On the other hand, we extract regions of homogeneity which are obtained by using two different image segmentation methods: We compare the well-known Mean-Shift-segmentation [5] with our similarity-measure-segmentation. This new segmentation method allows the segmentation of nonconnected regions. It performs equally well or better than several other methods with respect to object recognition in our experiments. Next, we calculate local descriptors of regions of discontinuity and homogeneity. Having various descriptions of the content of an image allows us to combine various kinds of regions with various descriptions in one learning step. We use Boosting [12] as learning technique. Boosting is a technique for combining several weak classifiers into a final strong classifier. The weak classifiers are calculated on different weightings of the training examples. This is done to emphasize different aspects of the training set. Since any classification function can potentially serve as a weak classifier, we can use classifiers based on arbitrary and diverse sets of image features. A further advantage of Boosting is that weak classifiers are calculated when needed instead of calculating unnecessary hypotheses a priori. The result of the training procedure is saved as the final hypothesis.

Existing data sets for object recognition used by other research groups (e.g., [8], [1]) show the objects with just small variations in scale and objects are generally viewed at similar poses. To be comparable with other state-of-the-art approaches, we also carried out experiments on the well-known Caltech[3] and the University of Illinois[4] databases. Fig. 2 shows some examples of the Caltech database of the categories cars (rear), motorbikes, and airplanes. On such databases, other previous approaches work well, because of the prominent objects. However, we require far more complex images to be able to demonstrate the advantages of our approach. The objects should be shown with high variation of their location in the image, at different scales, viewed from several positions. Additionally, the images should contain high background clutter. Therefore, we had to build up our own more complex database. This database[5] (further on termed GRAZ-01) that was used in [30], contains 450 images of category person (P), 350 of category bike (B), and 250 of category "counter-class" (N, meaning it contains no bikes and no persons). Fig. 3 shows some example images of each category.

Based on our localization results (see Section 7.3), which reveal that certain methods tend to emphasize context (i.e., the final classifier contains many background features), we have set up a second database (see footnote 5, further on termed GRAZ-02). This database has been carefully balanced with respect to background, such that similar backgrounds occur for all categories. Furthermore, we increased the complexity of the object appearances and added a third category of images. This challenging database contains 311 images of category person (P), 365 of category bike (B), 420 of category car (C), and 380 of a counter-class (N, meaning it contains no bikes, no persons, and no cars). Fig. 4 shows some example images. Our approach should cope with a high amount of occlusion and with significant scale changes. The images include all these difficulties with occlusions up to 50 percent. Also, the scale of the objects varies around 5 times of their average size.

Regarding different region detection and description techniques shown in Fig. 1, we experimentally evaluate two kinds of methods. First, we perform various experiments for one region extraction with one kind of local description technique. We do not experiment with all possible combinations, but we focus on methods with high performance based on results reported in [29] and [30]. The second method is the combination of various kinds of region detections with

[2] Note that we do not use color information in this work. This might be a possible area of future improvement.

[3] Available at http://www.vision.caltech.edu/html-files/archive.html.

[4] Available at http://l2r.cs.uiuc.edu/~cogcomp/index_research.html.

[5] Available at http://www.emt.tugraz.at/~pinz/data/.

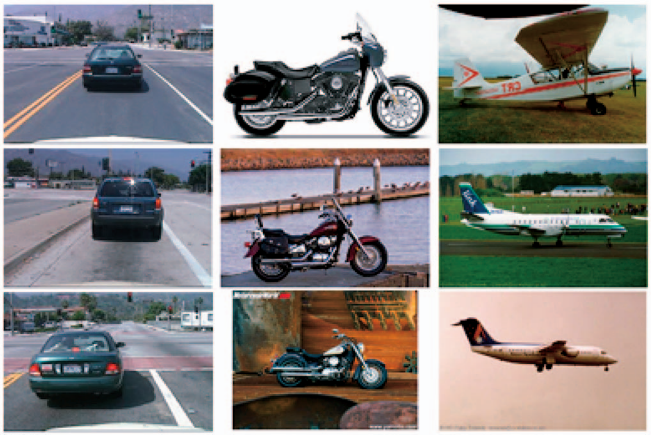


Fig. 2. Sample images from the Caltech database, categories cars (rear), motorbikes, and airplanes, used e.g., by [8].

different description techniques in one learning step (using the "combination" module shown in Fig. 1).[6]

The performance is measured by the commonly used receiver-operating-characteristic(ROC)-equal error rate (for details, see [1]). In some cases, we also report the ROC-AuC rate (area under ROC curve).

4 Region Detection

Using all the information of the whole image leads to a very high-computational complexity of the learning procedure. Therefore, a reduction of information is necessary. This can be achieved using salient information extraction techniques. But, we also want to be capable of learning many object categories without restrictions to shape or appearance of the objects. Each category might be characterized by different descriptors. For some objects, salient point techniques might be the best way to extract their essential information. For other objects, segments might be more relevant for recognition. Hence, an approach for generic object recognition would be limited if the images were described by just one method. While all existing approaches (e.g., [9], [1], [37]) use just one kind of description method for local image regions, we combine multiple information extraction methods. This should capture the essential characteristics of various object categories (e.g., persons, cars, etc). Complementing our approach, Viola et al. [39] use motion information as a second source of information, whereas we use various techniques to describe image intensity information. The increased complexity is justified by the gain of generalization in our approach.

There are two main branches of information extraction in our framework. The first one is to select regions of discontinuity. We use various well-known interest point extraction techniques and simply crop out a region (of a scale dependent size) around each point. The other branch is the extraction of regions of homogeneity. This means information reduction by a representation through image segments. We use our new similarity-measure-segmentation and compare it with Mean-Shift-segmentation.

4.1 Regions of Discontinuity

As mentioned, regions of discontinuity are regions around interest points. There is a variety of work on interest point detection at fixed (e.g., [17], [18], [36], [42]) and at varying scales (e.g., [22], [26], [27]). Based on the evaluation of interest point detectors by Schmid et al. [31], we decided to use the scale invariant Harris-Laplace detector [26] and the affine invariant interest point detector [27], both by Mikolajczyk and Schmid. In addition, we use Lowe's DoG (difference of Gaussian) keypoint detector [23] which is strongly related to SIFTs as local descriptors. As these techniques are state-of-the-art, we do not describe them in detail here. The interested reader is referred to the given references. We used the same parameter settings as the authors in their experiments. For the Harris-Laplace detector and the affine invariant interest point detector, we normalized the regions around the interest points to square patches with a side length of $w = 6 \cdot \sigma_I$ (ajar to the value used by Mikolajczyk and Schmid in [27]). Our framework calculates local descriptors from square patches of size $l \times l$. Scale normalization is achieved by smoothing and subsampling in cases of $l < w$ and by linear interpolation otherwise. For illumination normalization, we use homomorphic filtering (see e.g., [14], chapter 4.5). For DoGs, we used the binary of Lowe that already exports the local descriptors (SIFTs) of a circular region with a radius of eight pixels around the detected interest points.

4.2 Regions of Homogeneity

Regions of homogeneity can either be regions with a limited difference of intensity values or regions with homogeneous texture. These homogeneous regions are found with region-based segmentation algorithms. There is an extensive body of literature that deals with region-based segmentation algorithms and their applications. Many of them (e.g., [4] and [35]) are trying to split images into background and prominent foreground objects. Barnard et al. [2] use these segmentation methods to learn object categories. The advantage of this approach is the reduced complexity because there are only a few regions in each training image. The drawback is the difficulty to describe large and complex regions. Therefore, we prefer to use algorithms, which deliver more and smaller regions. These regions can be sufficiently well represented by simple descriptors (see Section 5).

We also have developed a new algorithm—"Similarity-Measure-Segmentation" (first presented in [13])—which is subsequently described in detail. We compare its performance for object categorization with the well-known Mean-Shift algorithm by Comaniciu and Meer [5]. In our framework, we use the code from "The Robust Image Understanding Laboratory."[7] Note that we just briefly compare the qualitative results of these segmentation methods. Then, we rather focus on their performance within our recognition framework.

4.2.1 Similarity-Measure-Segmentation

Similar to other segmentation algorithms (see [4] and [35]), we calculate several features for each pixel of the image, in a first processing step. But, in contrast to others, we use a similarity measure SM to describe pixel similarity for segmentation purpose:

$$SM = \frac{\sum_{i=1}^{n} a_i e^{-\frac{SC_i}{2\pi\sigma_i}}}{\sum_{i=1}^{n} a_i} \qquad 0 < SM \leq 1. \tag{1}$$

[6] Note that, even if the combination seems more interesting, we also want to compare the performance of the various methods separately.

[7] Available at http://www.caip.rutgers.edu/riul/research/code.html.

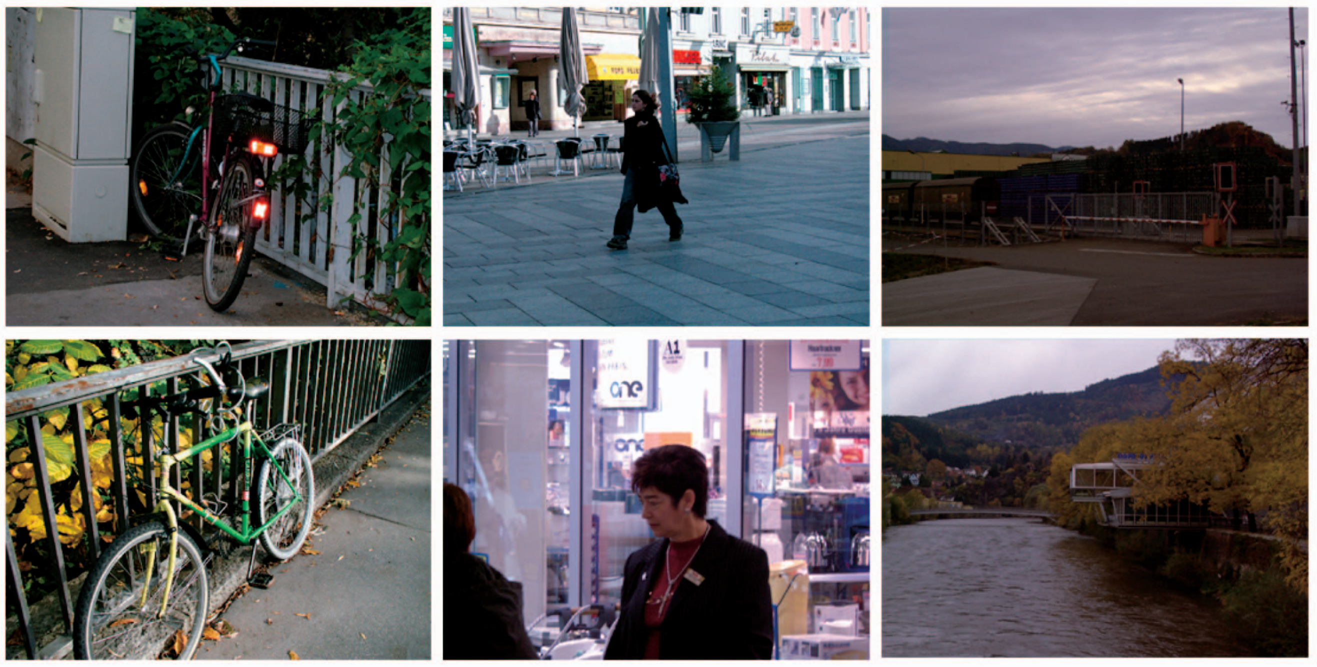


Fig. 3. Some example images from our database GRAZ-01. The first column shows examples of the category bikes (B). In the second column, there are images of the category person (P). The right-most column shows images of the counter-class (N). All these images were correctly classified using our approach (for details, see Section 7).

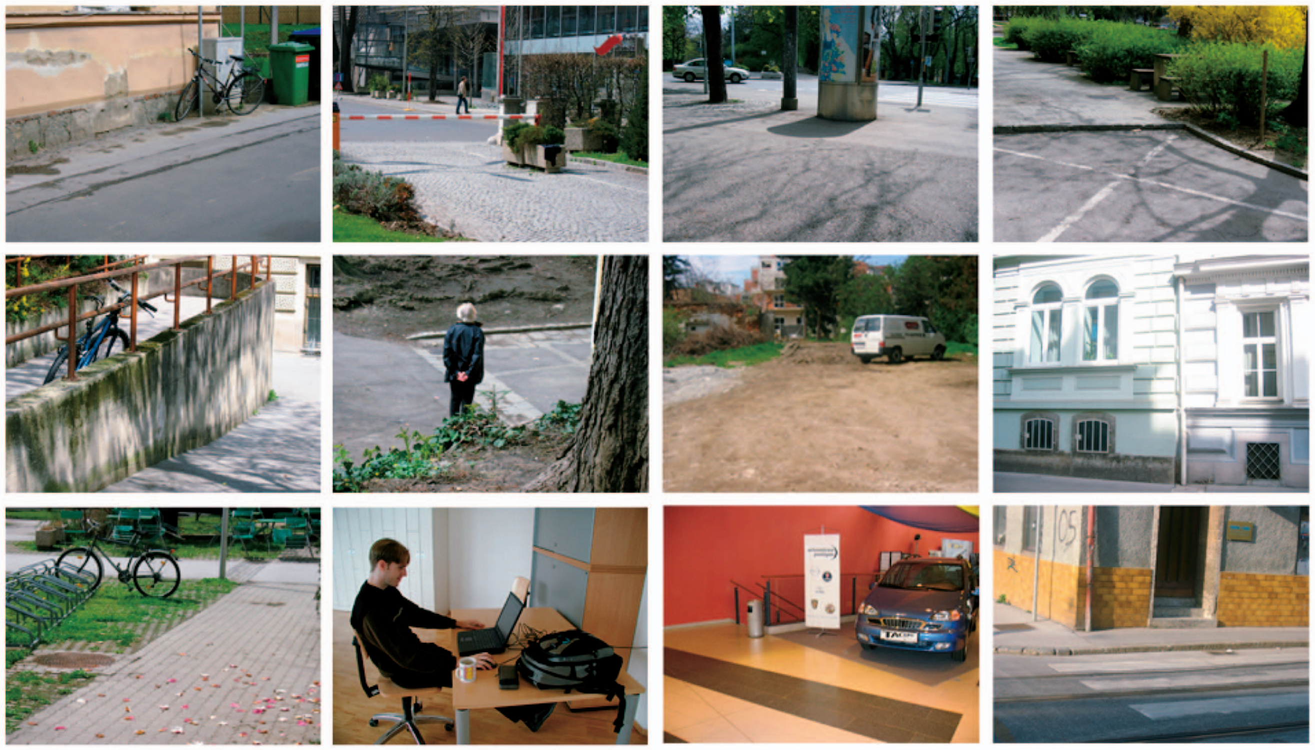


Fig. 4. Some example images from our database GRAZ-02. The first column shows examples of the category bikes (B). In the second column, there are images of the category person (P) followed by images of the category cars (C) in the third column. The right-most column shows some images of the counter-class (N). The complexity increased compared with the database GRAZ-01. Also, the appearances of the background of the images (category and counter-class) are rather balanced. All these images were correctly classified using our approach (for details, see Section 7).

This similarity is used to split images into regions. SC_i defines an element of the similarity-criteria vector SC. This can be seen as the distance of two pixels corresponding to a defined pixel feature. The parameter a_i can be set between 0 and 1 to change the weight of the similarity-criterion SC_i and σ_i is used to change the sensitivity. For example, on images with a small intensity variation, a small σ_i is used to enhance the sensitivity of the intensity similarity-criterion.

Our Similarity-Measure Grouping algorithm consists of the following steps:

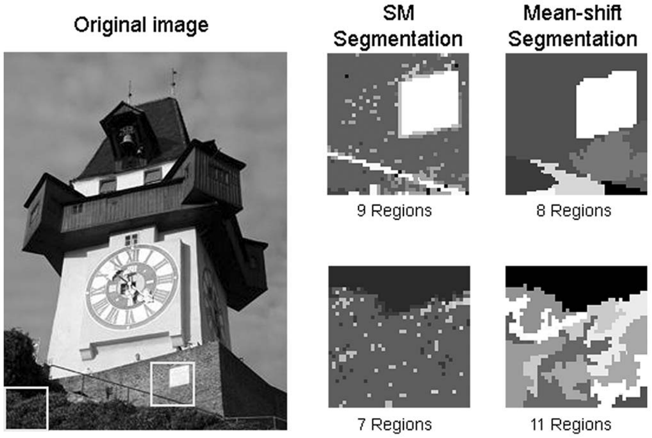


Fig. 5. Two detail views of the "Grazer Clocktower" segmented with Similarity-Measure segmentation (images in the middle) and Mean-Shift segmentation (images on the right).

1. Take any unlabeled pixel in an image, define a new region R_j, and label this pixel with RL_j.
2. Calculate the similarity measure to all other unlabeled pixels in the neighborhood, defined by a radius r.
3. Each pixel that has a similarity above a threshold t $(0 < t \leq 1)$ is also labeled with RL_j. Go back to Step 2 for each newly labeled pixel.
4. If there aren't any newly labeled pixels, start again with Step 1, until all pixels have a region number RL_k.
5. Search all regions smaller than a minimum value reg_{min} and merge each region with the nearest region larger than reg_{min} (same process as for Mean-Shift segmentation [5]).

The radius r can be varied between 1 and r_{max}. The maximum radius r_{max} depends on a positional sensitivity σ_x and on a threshold t:

$$r_{max} = \ln\left(\frac{t}{\sum_{i=1}^{n} a_i}(n-1)\right)(-2\pi\sigma_x). \qquad (2)$$

If we use $r = 1$, we have a region growing algorithm using the similarity-measure as homogeneity function. If we set the radius $r > 1$ (generally, $r = r_{max}$), we have a new segmentation method, that delivers not connected "regions" R_j. While this is in contradiction to the classical definition of segmentation, treating these R_j as entities for the subsequent learning process has shown recognition results, which are superior to results based on connected regions. We consider this new way of looking at disconnected segments a possibility to aggregate larger entities which are well-suited to describe local homogeneities. These descriptions maintain salient local information and suppress spurious information which would lead to oversegmentation in other segmentation algorithms.

Fig. 5 shows two detail views segmented with similarity-measure and with Mean-Shift segmentation. The first example shows a rail that disappears with Mean-Shift segmentation but is maintained with similarity-measure segmentation. The rail is disconnected because of some similarities between rail parts and the background, in both algorithms. The Mean-Shift algorithm merges the maintaining rail parts with the background considering its two constraints that regions have to be connected and must be larger than reg_{min}. The Similarity-Measure algorithm treats the disconnected parts as one region, which is larger than reg_{min}. The second example shows a part of a bush. The bush is split into 11 small regions with Mean-Shift segmentation. Similarity-Measure segmentation leads to five disconnected regions surrounded by two large regions. This behavior is desirable for our purpose because it turns out that a representation by not connected regions leads to a better performance of our categorization approach.

5 Local Description

For the learning step, each region has to be represented by some local descriptors. We use different description methods for the two region types.

For regions of discontinuity, local descriptors have been researched quite well (e.g. [11], [23], [32], [15]). We selected four local descriptors for these regions, namely: Subsampled gray values, basic intensity moments, moment invariants [15], and SIFTs [23]. This choice was partly based on the performance evaluation of local descriptors done by Mikolajczyk and Schmid [28]. For regions of homogeneity, we chose two description techniques: intensity distributions [16] and invariant moments [25]. The remaining part of this section gives a very brief explanation of these techniques.

Our first descriptor is simply a vector of all pixels in a patch subsampled by two. The dimension of this vector is $\frac{l^2}{4}$, which is rather high and increases computational complexity. As a second descriptor, we use intensity moments $M^a_{I_{pq}} = \iint_\omega i(x,y)^a x^p y^q \, dx \, dy$ with a as the degree and $p+q$ as the order, up to degree 2 and order 2. Without using the moments of degree 0, we get a feature vector of dimension 10. This reduces the computational costs dramatically.

According to [15], we selected first and second order moment invariants. We chose four first order affine and photometric invariants. Additionally, we took all five second order invariants described in [15]. Since the invariants require two contours, the whole region (square patch) is taken as one contour and rectangles corresponding to one half of the patch are used as a second contour. All four possibilities of the second contour are calculated and used to obtain the invariants. The dimension of the moment invariants description vector is 9.[8]

As shown in [23], the description of the patches with SIFTs is done by multiple representations in various orientation planes. A local descriptor with a dimension of 128 is obtained.

The last two methods are used to represent regions of homogeneity. The first one describes the intensity values and their distribution in a region. It contains their mean, variance, coefficient of variation, smoothness, skewness, kurtosis, and the gray value energy (for details, see [16]). The second one contains invariant moments (see [25]), which are invariant with respect to scaling, rotation, and translation. They are calculated from basic moments of inertia. Using basic moments of order up to three results in seven invariant moments for this description method.

Table 1 gives an overview of the various description methods in our framework and their dimension.

6 Learning Model

Our learning model is based on the AdaBoost algorithm [12]. This algorithm was adapted by adding the possibility

[8] Note that, we skip description vectors here, which have more than five entries equal to zero. This improved our results using moment invariants.

TABLE 1
An Overview of the Description Methods in Our Framework and Their Dimension (for a Region Size of $l = 16 \times 16\ pixels$)

-	Regions of discontinuity				Regions of homogeneity	
Method	Subsampled grayval.	Basic moments	Moment Invariants	SIFTs	Intensity distribution	Invariant moments
Dimension	64	10	9	128	7	7

```
Input: Training images $(I_1, \ell_1), \dots, (I_m, \ell_m)$.
Initialization: Set the weights $w_1 = \cdots = w_m = 1$.

For $t = 1, ..., T$
    1) Get a weak hypothesis $h_t$ in respect to the weights $w_1, \dots, w_m$ from the Weak-Hypotheses-Finder.
    2) Calculate $\varepsilon = \frac{\sum_{k=1, h_t(I_k) \neq \ell_k}^{m} w_k}{\sum_{k=1}^{m} w_k}$.
    3) Choose $\beta_t = \begin{cases} \sqrt{\frac{1-\varepsilon}{\varepsilon}} \cdot \eta & \text{if } \ell_k = +1 \text{ and } \ell_k \neq h_t(I_k). \\ \sqrt{\frac{1-\varepsilon}{\varepsilon}} & \text{else} \end{cases}$
    4) Update $w_k \leftarrow w_k \cdot \beta^{-\ell_k \cdot h_t(I_k)}$ for $k = 1, \dots, m$.
Output the final hypothesis (classifier):
    $H(I) = \begin{cases} +1 & \text{if } \sum_{t=1}^{T} (\ln \beta_t) h_t(I) \geq th_{Ada}, \\ -1 & \text{else.} \end{cases}$
```

Fig. 6. Modified AdaBoost algorithm [12] for object categorization tasks.

of putting different weights on positive and negative training images. We set up a new weak-hypotheses-finder that selects the most discriminant description vector in each iteration of the AdaBoost algorithm. This weak-hypotheses-finder is extended to be capable of using various description methods in one learning step.

We need to learn a classifier for recognizing objects of a certain category in still images. For this purpose, the learning algorithm delivers a classifier that predicts whether a given image contains an object from this category or not. As training data, labeled images $(I_1, \ell_1), \dots, (I_m, \ell_m)$ are provided for the learning algorithm where $\ell_k = +1$ if I_k contains a relevant object and $\ell_k = -1$ if I_k contains no relevant object. The learning algorithm delivers a function $H : I \mapsto \hat{\ell}$ which predicts the label of image I.

6.1 AdaBoost

To calculate this classification function H, we use an adaptation of the classical AdaBoost algorithm [12]. AdaBoost puts weights w_k on the training images and requires the construction of a weak hypothesis h which has some discriminative power with respect to these weights, i.e.,

$$\sum_{k=1, h(I_k)=\ell_k}^{m} w_k > \sum_{k=1, h(I_k) \neq \ell_k}^{m} w_k \qquad (3)$$

such that more images are correctly classified than misclassified, relative to the weights w_k. Such a hypothesis is called weak since it needs to satisfy only this very weak requirement. The process of putting weights and constructing a weak hypothesis is iterated for several rounds $t = 1, \dots, T$, and the weak hypotheses h_t of each round are combined into the final hypothesis H (for details, see Fig. 6). We use a threshold th_{Ada} (in [12], the authors use a signum function which means $th_{Ada} = 0$) to get the final classification result. To generate various points on the ROC curve, one can train a classifier and then use varying values for the threshold th_{Ada}.

In each round t, the weight w_k is decreased if the prediction for I_k was correct ($h_t(I_k) = \ell_k$) and increased if the prediction was incorrect. Different to the standard AdaBoost algorithm, we vary the calculation of the factor β_t which AdaBoost uses for its weight update after each iteration. We add a possibility to trade off precision and recall. We set

$$\beta_t = \begin{cases} \sqrt{\frac{1-\varepsilon}{\varepsilon}} \cdot \eta & \text{if } \ell_k = +1 \text{ and } \ell_k \neq h_t(I_k) \\ \sqrt{\frac{1-\varepsilon}{\varepsilon}} & \text{else} \end{cases}$$

with ε being the error of the weak hypothesis in this round and η as an additional weight factor to control the update of falsely classified positive examples.

Here, two general comments are in place. First, it is intuitively quite clear that weak hypotheses with high discriminative power—with a large difference of the sums in (3)—are preferable and, indeed, this is shown in the convergence proof of AdaBoost [12]. Second, the adaptation of the weights w_k in each round performs some sort of adaptive decorrelation of the weak hypotheses: If an image was correctly classified in round t, then its weight is decreased and less emphasis is put on this image in the next round. As a result, this yields quite different hypotheses h_t and h_{t+1}[9] and it can be expected that the first few weak hypotheses characterize the object category under consideration quite well. This is particularly interesting when a sparse representation of the object category is needed.

[9] In fact, AdaBoost sets the weights in such a way that h_t is *not* discriminative with respect to the *new* weights. Thus, h_{t+1} is in some sense oblivious to the predictions of h_t.

Input: Labeled representations $(\mathcal{R}(I_k), \ell_k)$, $k = 1, \ldots, m$, $\mathcal{R}(I_k) \triangleq \{(\tau_{k,f}, v_{k,f}) | f = 1, \ldots, F_k\}, (w_k)$.

(1): Distance functions: Let $d_\tau(\cdot, \cdot)$ be the distance in respect to the description vectors of type τ in the training images.

(2): Minimal distance matrix: For all description vectors $(\tau_{k,f}, v_{k,f})$ and all images I_j calculate the minimal distance between $v_{k,f}$ and description vectors in I_j,

$$d_{k,f,j} = \min_{1 \le g \le F_j : \tau_{j,g} = \tau_{k,f}} d_{\tau_{k,f}}(v_{k,f}, v_{j,g}) \ .$$

(3): Sorting: For each k, f let $\pi_{k,f}(1), \ldots, \pi_{k,f}(m)$ be a permutation such that

$$d_{k,f,\pi_{k,f}(1)} \le \cdots \le d_{k,f,\pi_{k,f}(m)} \ .$$

(4): Select best weak hypothesis (Scanline): For all description vectors $(\tau_{k,f}, v_{k,f})$ calculate over all images I_j

$$\max_s \sum_{j=1}^{s} w_{\pi_{k,f}(j)} \ell_{\pi_{k,f}(j)} \ .$$

and select the description vector $(\tau_{k,f}, v_{k,f})$ where the maximum is achieved.

(5): Select threshold θ: With the position s where the scanline reached a maximum sum the threshold θ is set to

$$\theta = \frac{d_{k,f,\pi_{k,f}(s)} + d_{k,f,\pi_{k,f}(s+1)}}{2} \ .$$ [a]

[a] This does not necessarily minimize the error, if $d_{k,f,\pi_{k,f}(s)} = d_{k,f,\pi_{k,f}(s+1)}$.

Fig. 7. Weak-hypotheses-finder using various description methods at a time.

Obviously, AdaBoost is a very general learning technique to obtain classification functions. To adapt for a specific application, suitable weak hypotheses have to be constructed. For the purpose of object recognition, we need to extract suitable description vectors from images and use these descriptors to construct the weak hypotheses. Since AdaBoost is a general learning technique, we are free to choose any type of description method we like, as long as we are able to provide an effective weak-hypotheses-finder which returns discriminative weak hypotheses based on this set of descriptors. The chosen description vectors should be able to represent the content of images, at least with respect to the object category under consideration.

Since we can choose several types of description vectors, we represent an image I by a set of pairs $\mathcal{R}(I) = \{(\tau, v)\}$, where τ denotes the type of a descriptor and v denotes a value of this descriptor, typically, a vector of reals. Then, for AdaBoost, a weak hypothesis is constructed from the representations $\mathcal{R}(I_k)$, labels ℓ_k, and weights w_k of the training images.

6.2 Weak-Hypotheses-Finder

Using one type of description vector at a time is the basic functionality of our learning algorithm. But, it is also possible to use multiple description methods in one learning procedure. Then, the challenge of the learning algorithm is not only the selection of the most discriminant description vector with respect to the current weighting but also the choice of a description type τ.

An image I_k is represented by a list of descriptors $(\tau_{k,f}, v_{k,f})$, $f = 1, \ldots, F_k$. The weak hypotheses for AdaBoost are calculated from these descriptors. Fig. 7 shows the weak-hypotheses-finder using multiple description methods. For object recognition, we have chosen weak hypotheses which indicate if certain description vectors appear in images. That is, a weak hypothesis h has to select a feature type τ and its value v and a similarity threshold θ. The threshold θ decides if an image contains a description vector $v_{k,f}$ that is sufficiently similar to v. The similarity between $v_{k,f}$ and v is calculated by the Mahalanobis distance for moment invariants, basic intensity moments and the descriptors for the regions of homogeneity. Euclidean distance is used for the SIFTs and the subsampled gray values due to the high dimension of the feature space. The weak-hypotheses-finder (Fig. 7, Step 4) searches for the optimal weak hypotheses—given labeled representations of the training images $(\mathcal{R}(I_1), \ell_1), \ldots, (\mathcal{R}(I_m), \ell_m)$ and their weights $w_1, \ldots, w_m$ calculated by AdaBoost—among all possible description vectors and corresponding thresholds. Our learning algorithm is simplified if various description methods τ are used separately.

The main computational burden is the calculation of the distances between $v_{k,f}$ and $v_{j,g}$ (see Fig. 7, Step 2) because they both range over all description vectors that appear in the training images. We arrange the minimum distances from each description vector to each image in a matrix, where we sort the distances in each column. Given these sorted distances, which can be calculated prior to Boosting, the remaining calculations are relatively inexpensive. In detail, we first calculate the optimal threshold for the description vector $v_{k,f}$ in time $O(m)$ by scanning

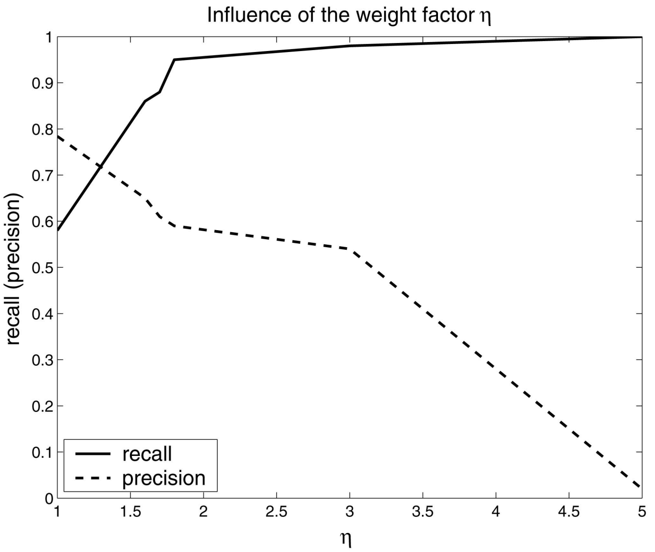


Fig. 8. The diagram shows the influence of the additional factor η for the weight-update of incorrectly positive classified examples. The recall increases faster than the precision drops until a factor of 1.8 (for the GRAZ-01 data set with affine invariant regions and moment invariants). The optimal value of this factor varies on different data sets.

through the weights $w_1, \ldots, w_m$ in the order of the distances $d_{k,f,j}$. Subsequently, we search over all description vectors. This calculation of the optimal weak hypothesis takes $O(F \cdot m)$ time, with F being the average number of features in an image.

To give an example of the total computation times, we use a data set of 150 positive and 150 negative images. Each image has an average number of approximately 400 description vectors. After preprocessing and using SIFTs, one iteration of Boosting requires about 10 seconds computation time on a P4 (2.4GHz PC). Obviously, the computational complexity is increasing with every additional kind of feature used.

7 Experiments and Results

The experimental evaluation is split into three parts. The first part (Section 7.1) specifies the parameter settings. Our classification results are discussed in detail in Section 7.2, showing that this approach clearly outperforms current state-of-the-art techniques. We investigate the performance using various features separately. The benefits of using multiple features in one learning procedure are also pointed out there. Section 7.3 presents a qualitative evaluation of localization performance. It shows the distribution of learned information that is directly related with the object and the learned contextual information.

7.1 Parameter Setting

The results were obtained using the same set of parameters for each experiment. For the regions of discontinuity (scale and affine invariant interest point detector), we used a threshold of cornerness $th = 30,000$ to reduce the number of salient points. Also, the points with the smallest characteristic scale were skipped (the neglectable influence of these points was shown in [29]). The side of the squared region size around the scaled and the affine interest points was normalized to $l = 16$ pixels. Vector quantization was used to reduce the number of interest points obtained with the difference of Gaussian (DoG) point detector [23]. Initially, we took all points into account but then we clustered the SIFT description vectors of each image. As a clustering algorithm, we used "k-means." The number of cluster centers cl was set to 100 (for the experiments on the GRAZ-02 database, we used $cl = 300$[10]) using a maximum number of 40 rounds in the k-means. For the extraction of the regions of homogeneity, we used a minimum region size $reg_{min} = 50$ for Mean-Shift-segmentation and similarity-measure-segmentation. We used the standard parameter set of the available binary for Mean-Shift-segmentation. For the similarity-measure-segmentation, we used a combination of intensity, position, and high-pass. We introduce σ_c for the intensity, σ_x for the position, and σ_t for the high-pass similarity criteria. σ_c depends on the contrast of the image. It is proportional to the variance of the image σ_I^2. The exact parameters used were: $\sigma_c = \frac{\sigma_I^2}{128} \cdot 3$, $\sigma_x = 1.2$, $\sigma_t = 0.5$, and a threshold of $t = 0.83$. With these parameters, we obtain $r_{max} = 6$. The learning procedure was run using $T = 100$.

TABLE 2
Relative Error on Data Set Cars (Rear) (Caltech Database) and Bikes (of GRAZ-01 and GRAZ-02) for One Point on the ROC Curve ($\eta = 1.0$, with $th_{Ada} = 0$)

Cars(rear) (Caltech)

Method	$reg_{min} = 50$	$reg_{min} = 250$
Mean-Shift	15	18.3
Similarity-M.	8.3	11.7

Bikes (GRAZ-01)

Method	$reg_{min} = 50$	$reg_{min} = 250$
Mean-Shift	18.3	23.3
Similarity-M.	15	20

Bikes (GRAZ-02)

Method	$reg_{min} = 50$	$reg_{min} = 250$
Mean-Shift	26.0	25.0
Similarity-M.	25.6	25.3

We compare similarity-measure-segmentation with Mean-Shift-segmentation. We used two different minimum region sizes of $reg_{min} = 50$ and $reg_{min} = 250$. In all cases, except for category bikes of GRAZ-02 with $reg_{min} = 250$, categorization results are better for similarity-measure-segmentation.

Fig. 8 shows the influence of the additional weight factor on recall and precision. In this test on the bike category of the GRAZ-01 data set, with affine invariant interest point detection and moment invariants, the optimal value is at $\eta = 1.8$. Up to this η, the recall increases faster than the precision drops. This optimal point depends on the description type and the data set. For all other experiments, we generally set $\eta = 1.0$ (standard AdaBoost), because this significantly accelerates the learning. If $\eta \neq 1.0$, we mention it separately.

The power of our new similarity-measure-segmentation with respect to object categorization is shown in Table 2. It outperforms Mean-Shift-segmentation in all cases, except for category bikes of GRAZ-02 with $reg_{min} = 250$, where they performed nearly equal. Thus, for the remaining experiments, we focused on regions of homogeneity obtained by similarity-measure-segmentation.

[10] These numbers were experimentally evaluated and depend on the image complexity, for details, see [29].

TABLE 3
Shows the ROC-Equal Error Rates on the Caltech Database and on Cars Side from the University of Illinois

Dataset	(1)	(2)	[8]	Others
Motorbikes	94.3	92.2	92.5	94.0 [20], 88 [41], 93.2 [37]
Airplanes	97.5	88.9	90.2	-
Faces	100	93.5	96.4	93.5 [41]
Cars(side)	100	83.0	88.5	97.5 [20], 79 [1]
Cars(rear)	100	91.1	90.3	93.9 [20], 86.5 [41]

The results in the first column (1) are obtained using regions of homogeneity extracted with the similarity-measure-segmentation and the description method based on the intensity distribution (with $\eta = 1.4$). The second column (2) shows the results using the affine invariant interest point detection and Moment Invariants. The last two columns show results for comparison.

7.2 Classification Results

7.2.1 Reference Data Set

To be comparable with existing approaches, we first evaluated our method on the Caltech database and "cars side" from the University of Illinois. We took regions of homogeneity extracted with the similarity-measure-segmentation and the description method based on the intensity distributions. We trained this combination on 60 images containing the object as positive training images and 60 images from the counter-class as negative training images. The tests were carried out on 60 new images half belonging to the learned class and half to the counter-class.[11] The results are shown in the first column of Table 3. The second column shows the results of our approach obtained with regions of discontinuity extracted with the affine invariant interest point detector and moment invariants as description method. Here, we trained this combination on 100 images containing the object as positive training images and 100 images from the background set as negative training images. We took 100 test images half belonging to the category and half not. In the last two columns, we compare our results with other state-of-the-art approaches ([1], [41], [8], [20], [37]). This comparison shows that our best results are superior to the classification performances of all the other approaches mentioned in the table. Note that, in the case of cars (side), we compare ROC-equal-error rates with the RPC-equal-error rates of other approaches. The other approaches face the harder task of also detecting multiple objects in one image. Whereas our model free approach cannot detect multiple instances of an object category in an image, but just reliably classify the whole image. Especially similarity-measure-segmentation-based region detection yields a very significant improvement on this data set.

7.2.2 GRAZ-01 Data Set

TABLE 4
Comparison of ROC-Equal Error Rates (Eq.Err.) and ROC-AuC (Area under Curve) Rates on GRAZ-01 Achieved with Three Specific Combinations: Affine Invariant Interest Point Detection with Moment Invariants, DoG Keypoint Detection Combined with SIFT as Description Method, and Similarity-Measure-Segmentation (SM) Described by Intensity Distributions

Dataset	Moment Invariants		SIFTs		SM	
-	eq.err.	AuC	eq.err.	AuC	eq.err.	AuC
Bikes	73.5	76.5	78.0	86.5	83.5	89.6
Persons	63.0	68.7	76.5	80.8	56.5	59.1

Having demonstrated the good performance of our approach on reference data sets (Caltech, Illinois), we proceed with experiments on our own GRAZ-01 database. We first took 100 images from the category bike (or person) as positive training images and 100 images of the counter-class (N) as negative training set. For the tests, we used 100 new images half containing the object (bike or person) and half not containing the object (category N).[12] On this set of images, we performed three experiments: First, we used regions of discontinuity extracted with the affine invariant interest point detection combined with moment invariants as description method. In the second experiment, we used regions of discontinuity obtained with the DoG keypoint detector combined with the SIFT description method. The number of cluster centers of the k-means was set to 100 in this experiment. Finally, we carried out an experiment using regions of homogeneity with intensity distributions as description method. Table 4 shows the ROC-equal error rates of each experiment for the categories bike and person. Considering the complexity of the data the results are very good. The best classification is obtained using Similarity-Measure-Segmentation (SM) described by intensity distributions for category bike and with DoG points and SIFTs for persons. This result shows that each category of objects is best represented by a specific description method. Fig. 9 shows the recall-precision curves of these experiments.

All images presented previously in Fig. 3 were categorized correctly. Fig. 10 gives examples of incorrectly classified images. In both cases, the images of the counter-class result from an experiment where we trained the category bikes.

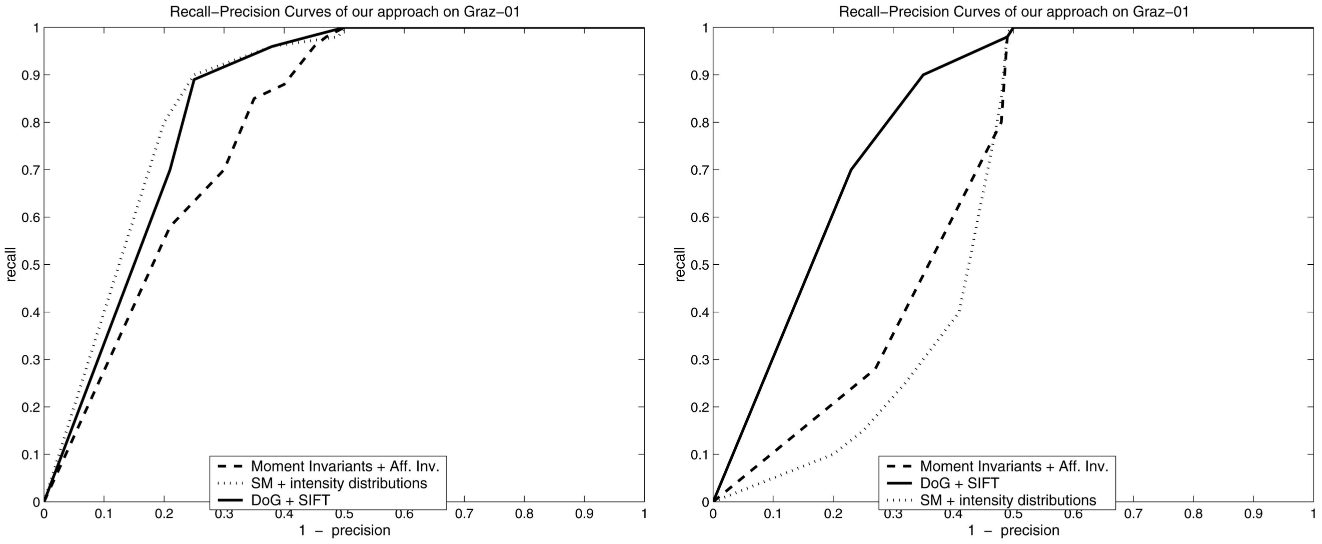
7.2.3 GRAZ-02 Data Set

After these experiments on the GRAZ-01 data set, we evaluated our approach using the GRAZ-02 data set. We took a training set consisting of 150 images of the object category as positive images and 150 of the counter-set as negative images. The tests were carried out on 150 images half belonging to the class and half not .[13] Fig. 11 shows the ROC curves of various specific combinations of region extractions and description types. Table 5 shows the resulting ROC-equal error rates. The affine invariant interest point detection with moment invariants or basic moments as local descriptors performs best except for the category bikes where all combinations achieve good results.

[11] The images are chosen sequentially from the database. This means, e.g., for this experiment, we took the first 90 image from the images of an object class and took out every third image for the test set.

[12] The images are chosen sequentially from the database. This means, e.g., for this experiment, we took the first 150 image from the images of an object class and took out every third image for the test set.

[13] The images are again chosen sequentially from the database. Note that the number of training images increases with the complexity of the data. With fewer images, our approach would not be able to fetch the category relevant information.



(a)

(b)

Fig. 9. Shows the recall precision curves of our approach. We compare Moment Invariants and the affine invariant interest point detection, SIFTs and DoG interest point detection, and Similarity-Measure-Segmentation (SM) described by intensity distributions on the GRAZ-01 database. (a) Shows the results for category bike and (b) shows the recall-precision curves for the category person.

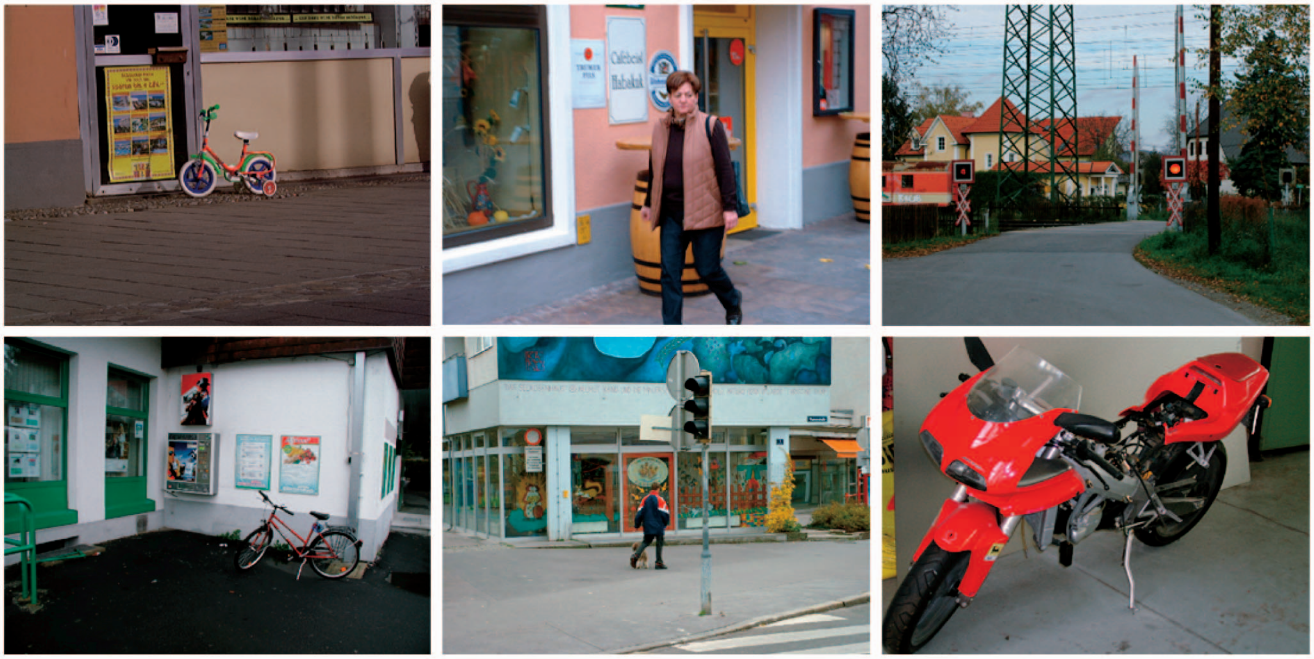
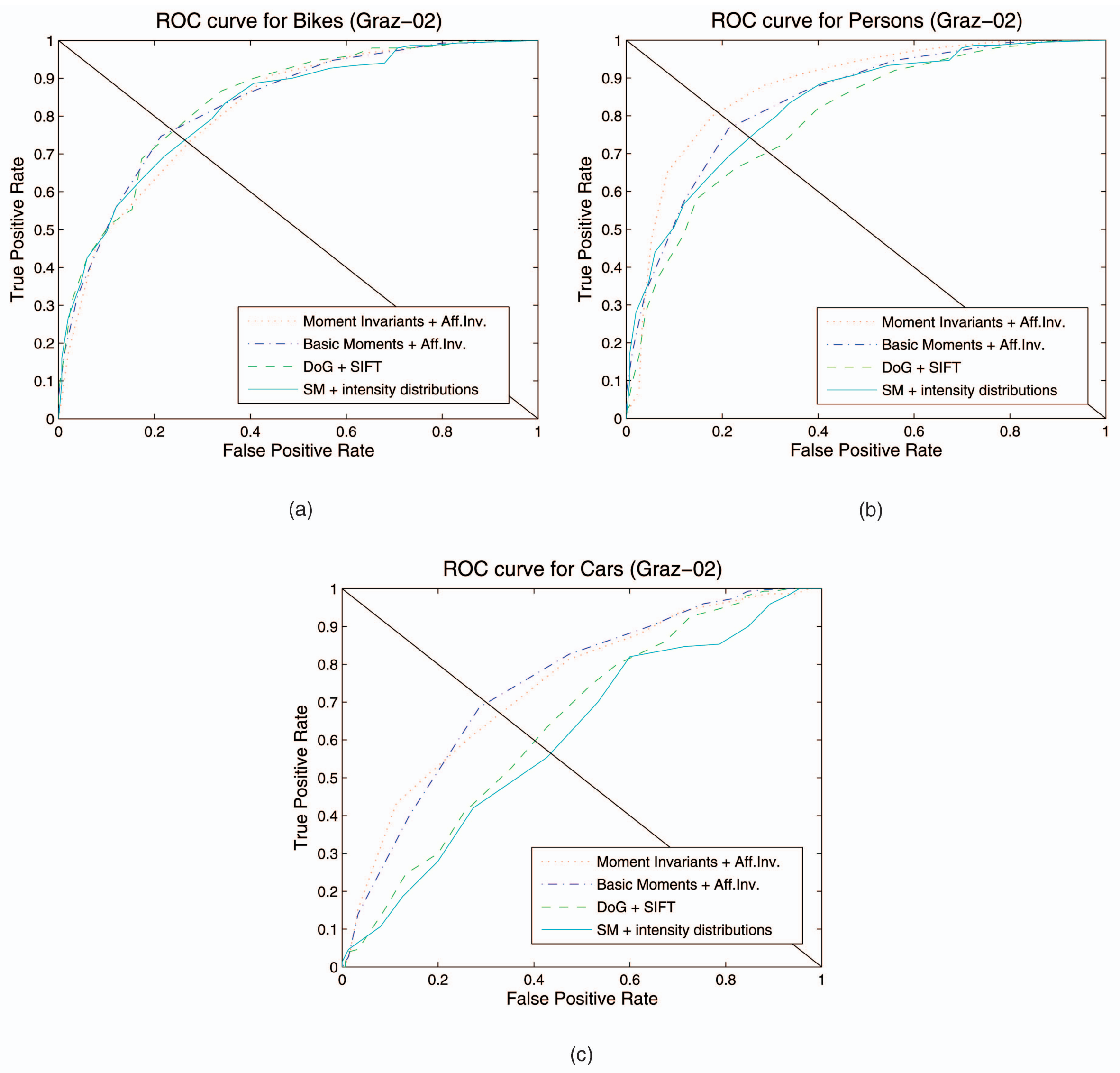


Fig. 10. Some example images from our database GRAZ-01 that were incorrectly classified in an average test case. The first column shows examples of the category bikes (B) classified as images not containing a bike. In the second column, there are images of the category person (P) classified as images not containing a person. The right-most column shows images of the counter-class-set (N) that were classified as bikes (B).

Again, all the images in Fig. 4 were categorized correctly while images in Fig. 12 represent examples, where the classification fails. One can see that the approach can handle quite huge scale variations (e.g., Fig. 4 second column). The system is even able to categorize an image where the object is occluded up to 50 percent (e.g., Fig. 4 second row, first column). However, it seems that too severe scale changes degrade the categorization performance (e.g., Fig. 12 first column, second row, or first row, third column).

This qualitative visual comparison of Figs. 3 and 10 with Figs. 4 and 12 immediately reveals the need of further explanation. Although the overall categorization results (regarding the highly complex data and the low supervision) are impressive, some difficult images are categorized correctly, while the method fails for other (sometimes “easier”) ones. What are the limitations of the approach? Why are certain images categorized incorrectly? Why do certain methods perform better than others? Especially, why is similarity-measure-segmentation a clear winner on the Caltech and Illinois data sets and on GRAZ-01 for the category bikes, still good on the GRAZ-02 bikes and persons, but quite poor on persons from GRAZ-01 and cars from GRAZ-02? We



(a)

(b)

(c)

Fig. 11. Shows the ROC curves of various specific combinations of region extractions and description methods on the three categories ((a) bikes, (b) persons, and (c) cars) of the GRAZ-02 data set.

try to answer some of these questions in Section 7.3 in the light of localization abilities of the various detectors.

7.2.4 Combination

Subsequently, we describe experiments performed by using more than one type of the various region extractions with a description method in one learning step. We evaluated three kinds of combinations.[14] In all cases, we use regions obtained with affine invariant interest point detection, described with moment invariants as the first method. We combine it with regions achieved through DoG keypoints described by SIFTs (see Table 6 (A)), regions extracted with the affine invariant interest point detector described with basic intensity moments (see Table 6 (B)) and regions of homogeneity obtained by the similarity-measure-segmentation and described with intensity distributions (see Table 6 (C)). While the results of the combinations show just slight enhancement over the individual best result, these experiments clearly show that the combination of several methods can perform significantly better than a certain individual method (cf. ROC-equal error rates of 81.2 versus 74.1 for persons). The main benefit is that a

TABLE 5
ROC-Equal-Error Rates of Various Specific Combinations of Region Extractions and Description Methods on the Three Categories of the GRAZ-02 Data Set

Dataset	Moment Invariants	Basic Moments	SIFTs	SM
Bikes	72.5	76.5	76.4	74.0
Persons	81.0	77.2	70.0	74.1
Cars	67.0	70.2	68.9	56.5

The first and the second column are obtained with the affine invariant interest point detection and moment invariants or basic intensity moments as local descriptor. The third row was achieved using DoG keypoint detection and SIFTs as description method using 300 cluster centers within the k-means clustering. The last column shows the results of experiments performed using similarity-measure-segmentation and description via intensity distributions.

[14] Combining more of our methods is just marginally improving the results.

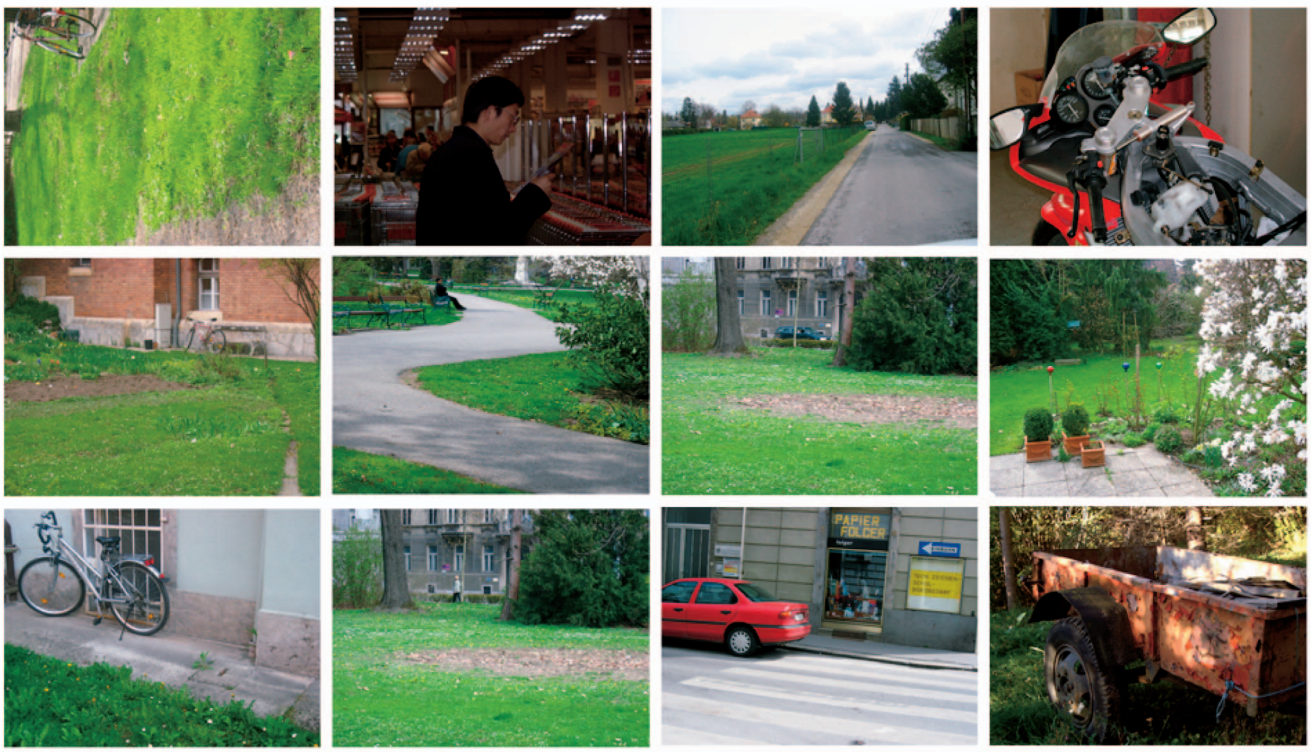


Fig. 12. Some example images from our database GRAZ-02 that were incorrectly classified in an average test case. The first column shows examples of the category bikes (B). In the second column, there are images of the category person (P) followed by images of the category cars (C) in the third column. All were classified as counter-class-images. The right-most column shows some images of the counter-class-set (N). These are examples that were classified as bikes (B).

use of the combination adds a higher reliability to a classifier. For some categories, one combination of a region extraction and a description method performs better than others. Using various specific combinations in one learning step ensures a final classifier that achieves better results than the best classifier used separately.

7.3 Localization Performance

To discuss the localization of the information learned by our approach, we first evaluated the experiments shown in the previous section with respect to the localization of the hypotheses. Taking a closer look at the regions of homogeneity that are learned to achieve the classification results of Table 3, we found out, that only 25 percent to 50 percent are located on the object. The remaining hypotheses do not learn the object category directly, but focus on contextual (background) information for this object category. Fig. 13 shows some examples of regions of homogeneity selected as weak hypotheses from the Caltech data set. The first row shows four hypotheses of the category plane. The first three regions are located on the plane whereas the last one is not. The second row shows four hypotheses from the final classifier of the category cars (rear). Again, the right-most hypothesis is not located on the object. If the object category of the data set has specific background appearances that do not occur in the images of the counter-class, it is in the nature of our learning approach to select also background information. Thus, this combination of object information and contextual information gives us a good classification performance. On the other hand, object localization is not

TABLE 6
This Table Shows the ROC-Equal Error Rates of Specific Combinations of Region Extractions and Description Methods Separated and Their Performance if They Are Combined in One Learning Step (on GRAZ-02)

Dataset	Mom. Inv.	method 2	combination
Cars	67.0	70.2 (A)	70.5
Bikes	72.5	76.4 (B)	77.8
Persons	81.0	74.1 (C)	81.2

The first value is always for the moment invariants. The second column shows the results of either basic intensity moments (A) or SIFTs (B) or regions of homogeneity described through intensity distributions (C). The last column shows the achieved performance using the combination of the two methods.

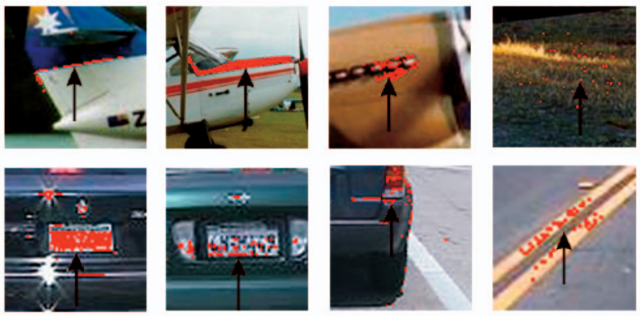


Fig. 13. Some examples of weak hypotheses of regions of homogeneity. The first row shows four hypotheses from the final classifier of the category airplane. In the second row, weak hypotheses of the category cars (rear) are shown.

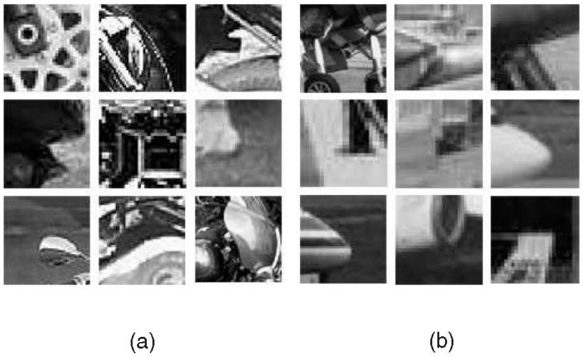


Fig. 14. (a) Shows nine examples of regions of discontinuity selected for the final classifier of the category motorbike. (b) Shows nine examples of regions of discontinuity selected for the final classifier of the category airplane.

straight forward if we use regions of homogeneity on images with specific background appearances.

Fig. 14a shows examples of regions of discontinuity learned as weak hypotheses for the category motorbikes. The final classifier was trained using affine invariant interest points and moment invariants as local description method. The regions shown are the raw image data cropped out around the interest point before any affine, illumination, and size normalization. Using the same settings, Fig. 14b shows weak hypotheses of the final classifier of the category airplanes. With this specific combination, we obtain 80 percent to 90 percent of the weak hypotheses located on the object. Even if this classifier is more related to the object (instead of containing contextual information), the classification result in Table 3 is lower compared to using regions of homogeneity.

TABLE 7
This Table Shows the Percentage of the Weak Hypotheses that Are Not Located on the Object

Dataset	Moment Invariants	Basic Moments	SIFTs	SM
Bikes	21	30	39	55
Persons	23	45	54	74
Cars	56	63	52	84

Here, we used the same combinations as in Table 5 for the GRAZ-02 data set.

Focusing on the percentage of contextual information that is learned, compared to the information directly related to the object, we took a closer look at the classifiers shown in Table 4 based on the GRAZ-01 data set. We observe an average of 60 percent of the weak hypotheses containing contextual information if we use similarity-measure-segmentation combined with intensity distributions. For DoG interest points described by SIFTs, 50 percent of the hypotheses contain contextual information. Using the affine invariant interest point detector with moment invariants or basic intensity moments decreases this percentage to 30 percent.

Table 7 shows the percentage of weak hypotheses of the final classifier for each category of GRAZ-02 that are not located on the object. Again, looking at Table 5 with respect to these localization performances shows that affine invariant interest point detection and moment invariants are most stable in the classification performance directly related to the object. Fig. 15 shows examples of weak hypotheses used for the final classifier of the category bike (GRAZ-02) with various description methods. It shows which information is

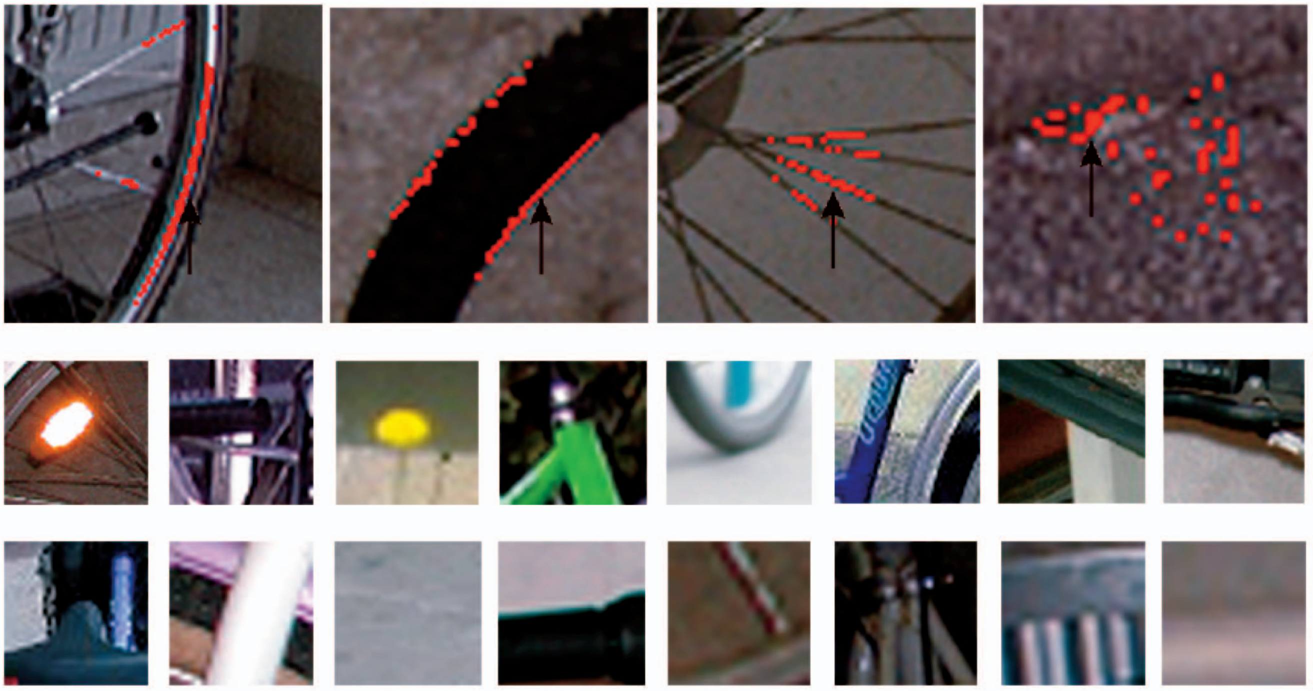


Fig. 15. Shows examples of weak hypotheses used for the final classifier of the category bike (GRAZ-02). The first row shows hypotheses based on the test with regions of homogeneity and intensity distributions. The middle row shows regions extracted with the affine invariant interest point detector and described by moment invariants. Examples of weak hypotheses obtained from the experiment with DoG keypoint detection and SIFTs are shown in the last row. These are the raw image patches before any normalization steps are carried out.

learned and how the learned classifier represents a category of objects. The hypotheses that contain background information (e.g., Fig. 15 first row, last column) are often also important for our classification. As most of the bikes occur associated with streets, weak hypotheses representing asphalt contain highly relevant contextual information.

In summary, these investigations lead to the following conclusions: The Caltech database shows the object of interest at very prominent scales, locations, and in very specific poses. The training data of the Illinois data set is even easier. While these constraints are significantly relaxed with the GRAZ-01 database, the counter-class images are quite different, which enables the algorithm to take background information (context) into account. It turns out, that homogeneity regions (similarity-measure-segmentation) and SIFTs tend to emphasize context more than other discontinuity-based region detectors. This is strongly supported by our results on the GRAZ-02 database, which is balanced with respect to the background (i.e., similar backgrounds for class and counter-class images).

8 Discussion and Outlook

We have presented a novel approach for the recognition of object categories in still images of high complexity. Our system uses several steps of region extraction and local description methods, which have been previously described, as well as a new segmentation technique, and succeeds on rather complex images with a lot of background structure. The only supervision we use are the image labels. We have set up new databases where objects are shown in substantially different poses and scales, and in many of the images the objects (bikes, persons, or cars) cover only a small portion of the whole image. We use Boosting as the underlying learning technique and combine it with a weak-hypotheses-finder. In addition to several other advantages of this approach, which have already been mentioned, we want to emphasize that this approach allows the formation of very diverse visual features into a final hypothesis. This use of several specific combinations of region extraction and description methods in one learning step makes a classifier more reliable over a whole range of different object categories. Furthermore, experimental comparison on the Caltech database shows that our approach performs better than state-of-the-art object categorization on simpler images. The new similarity-measure-segmentation turns out to be a powerful method to describe whole image contents.

We are currently investigating extensions of our approach in several directions. Maybe the most obvious one is the addition of more features to our image analysis. This includes not only other local descriptors, but also new regional features and geometric feature distributions. Also, the localization problem will be investigated in more detail. The different localization performances of various combinations in this framework might lead to the need of a loop within the learning procedure. There a first unsupervised localization step (with a technique that has good localization abilities) might be followed by the actual learning procedure which may contain several other methods. The new similarity-measure-segmentation should also be used for image retrieval in further experiments.

As a further step, we will use spatial relations between features to improve the accuracy of our object detector. Also, a loose geometrical model could improve our approach toward detecting multiple object instances in one image. To handle the complexity of many possible relations between features, we will use the features constructed in our current approach (with parameters set for high recall) as starting points. Boosting will again be the underlying method for learning object representations as spatial combinations of features. This will allow the construction of weak hypotheses for discriminative spatial relations.

Acknowledgments

This work was supported by the European project LAVA (IST-2001-34405), by the Austrian Science Foundation (FWF, project S9103-N04 and S9104-N04), and by the PASCAL Network of Excellence. The authors are grateful to David Lowe, Cordelia Schmid, and Krystian Mikolajczyk for providing the binaries of their detectors/descriptors, available on the Web.

References

[1] S. Agarwal and D. Roth, "Learning a Sparse Representation for Object Detection," *Proc. European Conf. Computer Vision*, pp. 113-130, 2002.

[2] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, and D. Forsyth, "The Effects of Segmentation and Feature Choice in a Translation Model of Object Recognition," *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 675-682, 2003.

[3] P. Carbonetto, N. de Freitas, and K. Barnard, "A Statistical Model for General Contextual Object Recognition," *Proc. European Conf. Computer Vision*, pp. 350-362, 2004.

[4] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1026-1038, Aug. 2002.

[5] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603-619, May 2002.

[6] G.Y. Dorko and C. Schmid, "Selection of Scale-Invariant Parts for Object Class Recognition," *Proc. Int'l Conf. Computer Vision*, pp. 634-640, 2003.

[7] P. Felzenszwalb and D. Huttenlocher, "Pictorial Structures for Object Recognition" *Int'l J. Computer Vision*, vol. 61, no. 1, pp. 55-79, 2004.

[8] R. Fergus, P. Perona, and A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning," *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 264-272, 2003.

[9] R. Fergus, P. Perona, and A. Zisserman, "A Visual Category Filter for Google Images," *Proc. European Conf. Computer Vision*, pp. 242-256, 2004.

[10] V. Ferrari, T. Tuytelaars, and L. Van Gool, "Simultaneous Object Recognition and Segmentation by Image Exploration," *Proc. European Conf. Computer Vision*, pp. 40-54, 2004.

[11] W. Freeman and E. Adelson, "The Design and Use of Steerable Filters," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 9, pp. 891-906 Sept. 1991.

[12] Y. Freund and R. Schapire, "A Decision Theoretic Generalisation of Online Learning," *Computer and System Sciences*, vol. 55, no. 1, pp. 119-139, 1997.

[13] M. Fussenegger, A. Opelt, A. Pinz, and P. Auer, "Object Recognition Using Segmentation for Feature Detection," *Proc. Int'l Conf. Pattern Recognition*, 2004.

[14] R.C. Gonzalez and R.E. Woods, *Digital Image Processing*. Addison-Wesley, 2001.

[15] L. Van Gool, T. Moons, and D. Ungureanu, "Affine/Photometric Invariants for Planar Intensity Patterns," *Proc. European Conf. Computer Vision*, pp. 642-651, 1996.

[16] R.M. Haralick, "Statistical and Structural Approaches to Texture," *Proc. IEEE*, vol. 67, pp. 786-804, 1979.

[17] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," *Proc. Fourth Alvey Vision Conf.*, pp. 189-192, 1988.

[18] R. Laganiere, "A Morphological Operator for Corner Detection," *Pattern Recognition,* vol. 31, no. 11, pp. 1643-1652, 1998.

[19] Y. LeCun, F.J. Huang, and L. Bottou, "Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting," *Proc. Conf. Computer Vision and Pattern Recognition,* 2004.

[20] B. Leibe, A. Leonardis, and B. Schiele, "Combined Object Categorization and Segmentation with an Implicit Shape Model," *Proc. European Conf. Computer Vision Workshop Statistical Learning in Computer Vision,* May 2004.

[21] T.K. Leung, M.C. Burl, and P. Perona, "Probabilistic Affine Invariants for Recognition," *Proc. Conf. Computer Vision and Pattern Recognition,* pp. 678-684, June 1998.

[22] T. Lindeberg, "Feature Detection with Automatic Scale Selection," *Int'l J. Computer Vision,* vol. 30, no. 2, pp. 79-116, 1998.

[23] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features," *Proc. Int'l Conf. Computer Vision,* pp. 1150-1157, 1999.

[24] W. Maass and M. Warmuth, "Efficient Learning with Virtual Threshold Gates," *Information and Computation,* vol. 141, no. 1, pp. 66-83, 1998.

[25] S. Maitra, "Moment Invariants," *Proc. IEEE,* pp. 679-699, 1979.

[26] K. Mikolajczyk and C. Schmid, "Indexing Based on Scale Invariant Interest Points," *Proc. Int'l Conf. Computer Vision,* pp. 525-531, 2001.

[27] K. Mikolajczyk and C. Schmid, "An Affine Invariant Interest Point Detector," *Proc. European Conf. Computer Vision,* pp. 128-142, 2002.

[28] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *Proc. Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 257-263, 2003.

[29] A. Opelt, "Feature Selection for Scaled Interest Points," master's thesis, Graz Univ. of Technology, 2003.

[30] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer, "Weak Hypotheses and Boosting for Generic Object Detection and Recognition," *Proc. European Conf. Computer Vision,* vol. 2, pp. 71-84, 2004.

[31] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of Interest Point Detectors," *Int'l J. Computer Vision,* vol. 37, no. 2, pp. 151-177, 2004.

[32] C. Schmid and R. Mohr, "Local Grayvalue Invariants for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 5, pp. 530-535, May 1997.

[33] H. Schneiderman and T. Kanade, "Object Detection Using the Statistics of Parts," *Int'l J. Computer Vision,* vol. 56, no. 3, pp. 151-177, 2004.

[34] A. Selinger and R.C. Nelson, "Improving Appearance-Based Object Recognition in Cluttered Background," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 1-8, 2000.

[35] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 8, pp. 888-905, Aug. 2000.

[36] E. Shilat, M. Werman, and Y. Gdalyahu, "Ridge's Corner Detection and Correspondence," *Proc. Computer Vision and Pattern Recognition,* pp. 976-981, 1997.

[37] J. Thureson and S. Carlsson, "Appearance Based Qualitative Image Description for Object Class Recognition," *Proc. European Conf. Computer Vision,* pp. 518-529, 2004.

[38] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proc. Conf. Computer Vision and Pattern Recognition,* pp. 511-518, 2001.

[39] P. Viola, M. Jones, and D. Snow, "Detecting Pedestrians Using Patterns of Motion and Appearance," *Proc. Int'l Conf. Computer Vision,* vol. 2, pp. 734-741, 2003.

[40] C. Wallraven, B. Caputo, and A. Graf, "Recognition with Local Features: The Kernel Recipe," *Proc. Int'l Conf. Computer Vision,* pp. 257-264, 2003.

[41] M. Weber, M. Welling, and P. Perona, "Unsupervised Learning of Models for Recognition," *Proc. European Conf. Computer Vision,* 2000.

[42] R.P. Wuertz and T. Lourens, "Corner Detection in Color Images by Multiscale Combination of End-Stopped Cortical Cells," *Proc. Int'l Conf. Artificial Neuronal Networks,* pp. 901-906, 1997.

Andreas Opelt received the MSc degree in computer science from Graz University of Technology in 2004. Currently, he is pursuing the PhD degree in the Department of Electrical Measurement and Measurement Signal Processing at Graz University of Technology. From October 2004 to June 2005, he served as an academic visitor in the Visual Geometry Group at the Oxford University. His research interests are focused on object recognition, machine learning, and cognitive vision.

Axel Pinz received the MSc degree in electrical engineering in 1983 and the PhD degree in computer science in 1988 from the Vienna University of Technology. In 1995, he received the habilitation degree in computer science from Graz University of Technology. He worked in high-level image analysis in remote sensing at the University of Natural Resources in Vienna (1983-1990, Institute of Surveying and Remote Sensing IVFL). From 1990 to 1994, he was an assistant professor in the Institute for Automation, Department of Pattern Recognition and Image Processing (PRIP), Vienna University of Technology. From 1994-1999, he was a visiting scientist in the Institute for Computer Graphics and Vision (ICG), Graz University of Technology, where he built up the Computer Vision Group of the Institute. In 1996 and 1997, he served as the academic head of the ICG, and from October 1997 to July 1999, he was a visiting professor in computer vision and computer graphics at Graz University of Technology, Austria. Since October 1999, he has been with the Institute of Electrical Measurement and Measurement Signal Processing (EMT), Graz University of Technology, Austria, where he is heading a research group that is focused on real-time measurement and object recognition. His main research interest is in high-level vision including spatio-temporal reasoning and tracking for object and scene recognition. He is a member of the IEEE.

Michael Fussenegger received the MSc degree in computer science from Graz University of Technology in 2003. Currently, he is pursuing the PhD degree in the Department of Electrical Measurement and Measurement Signal Processing at Graz University of Technology. From September 2004 to February 2005, he served as an academic visitor at the Odyssee-Lab at INRIA Sophia-Antipolis. His research interests are focused on segmentation and object recognition.

Peter Auer received the MSc and a PhD in mathematics from the Vienna University of Technology in 1987 and 1992. He has worked in probability theory with Professor Pal Revesz and in symbolic computation with Professor Alexander Leitsch (both at Vienna University of Technology, 1988-1991), and in machine learning with Professor Wolfgang Maass (Graz University of Technology, 1992-2002). He was a research scholar at the University of California, Santa Cruz, in 1995-1996. In 2003, he was appointed chair of information technology at the University of Leoben. He has authored and coauthored a significant number of refereed publications in scientific journals and conferences in the areas of probability theory, symbolic computation, and machine learning, and he is a member of the editorial board of *Machine Learning*. His current research interests include cognitive vision and machine learning, with an emphasis on autonomous and explorative learning methods.