Generic Object Recognition with Boosting

Andreas Opelt, Michael Fussenegger, Axel Pinz, Peter Auer

July 2004

Abstract

This paper presents a powerful framework for generic object recognition. Boosting is used as an underlying learning technique. For the first time a combination of various weak classifiers of different types of descriptors is used, which slightly increases the classification result but dramatically improves the stability of a classifier. Besides applying well known techniques to extract salient regions we also present a new segmentation method - "Similarity-Measure-Segmentation". This approach delivers segments, which can consist of several disconnected parts. This turns out to be a mighty description of local similarity. With regard to the task of object categorization, Similarity-Measure-Segmentation performs equal or better than current state-of-the-art segmentation techniques. In contrast to previous solutions we aim at handling of complex objects appearing in highly cluttered images. Therefore we have set up a database containing images with the required complexity. On these images we obtain very good classification results of up to 87% ROC-equal error rate. Focusing the performance on common databases for object recognition our approach outperforms all comparable solutions.

Index Terms

Adaboost, cognitive vision, object categorization, object localization.

I. Introduction

GENERIC object recognition is a difficult computer vision problem. Although various solutions exist for solving the recognition problem on common datasets, all these datasets have limitations. These limitations lower their complexity and make them differ from the idea of being really generic. Focusing on the generic case means: Considering datasets containing images with highly cluttered background. Also the images should show various arbitrary instances of the object category. This addresses two of the main problems faced by a visual categorization or recognition system: First, the difficulty of high intra class variability. And second, the need for generalization across variations in the appearances of objects of the same category. Different images of persons, for example, can change in various aspects. They can represent different persons. Also the appearance of each particular person will change because of various imaging parameters, like viewpoint, distance or illumination. A vision system faces the task of categorizing an unseen view of a known object. Or even categorize a new unknown instance of a certain category. A system must solve this problem using the learned experience to categorize the novel image. Another difficulty for learning an object category is the naturally high demand of supervision. It is hard for artificial vision systems to

Learning an object category from images with high background clutter, where the objects occur anywhere in the image is hard for artificial vision systems. Using a pile of images with persons and another pile without persons it should be obvious for the system what is the object of interest in these sets. This extraction of relevant information is naturally for humans but not for artificial vision systems. Facing these aspects, the human cognitive recognition system is so powerful, that comparable results in computer vision are hard to achieve. Humans learn somehow the wide spectrum of possible intra class variability

A. Opelt and A. Pinz are with the Institute of Electrical Measurement and Measurement Signal Processing, Graz University of Technology, Austria. Email: opelt@emt.tugraz.at, pinz@emt.tugraz.at

M. Fussenegger and P. Auer are with the Department of Mathematics and Information Technology, University Leoben, Austria. Email: fussenegger@emt.tugraz.at, auer@unileoben.ac.at

from experience. Hence object categorization is more complex for artificial systems than specific object recognition.

It is obvious that the less constraints an approach needs (on object appearance, background clutter, partial occlusion) to achieve acceptable results, the more generic it becomes. Each object category might have its specific description which represents it optimally. Consequently using one kind of model or representation for all object categories would hardly lead to a generic solution. This brings up the idea of combining various representations of objects to form a generic classifier for each category.

In this article we present a powerful framework for generic object recognition. It is based on an approach for learning objects in still images which allows the use of flexible and extendible sets of descriptions of objects and object categories. Objects should be recognized even if they appear at different scales, shown from different perspective views on highly textured backgrounds. The images are represented by salient regions described by various invariant description methods. Our new learning method is based on Boosting [12].

After a discussion of related work in section II, section III gives a detailed overview of our approach and explains our dataset and the difference to existing databases. In section IV 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 presented in section V followed by our general learning approach described in detail in section VI. The combination of various kinds of description vectors in the learning procedure is also shown in this section. Section VII describes our experimental setup, presents experimental results and compares them with other approaches for object recognition. Section VIII concludes with a discussion and an outlook on further extensions.

II. Related Work

Taking a closer look at the extensive body of literature on object recognition each approach has its specific 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], [2]). Others pre-segment the object manually (e.g. [7], [34]) 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 [35] 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. In our approach we experiment with much larger sets of features to be able to perform recognition of a wider class of objects on more complicated images. Schneiderman and Kanade [31] also 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, they 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 [2] 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. 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 which 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

[1]More efficient techniques for Winnow like using virtual threshold gates [21] do not improve the situation much.

when needed. To obtain their classification results, Agarwal and Roth crop out the object manually to reduce complexity.

Walraven et al. [36] 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 a generative probabilistic model for objects built as constellations of parts. Using an EM-type learning algorithm a very good recognition performance is achieved. They extend their constellation model in [9] to include heterogenous parts. The parts and their constellations can now be learned without supervision and from cluttered images. We use a model-free approach and propose Boosting as a very different learning algorithm compared with EM.

Another object recognition approach was introduced by Dorko and Schmid [7]. 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 pre-segmented. 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 non-cluttered representations of the object. Furthermore, this approach seems to be restricted to specific object recognition.

A new possibility of describing objects for categorization is introduced by Thureson and Carlsson in [34]. 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 representation images. Each new image is categorized for that region where the inner product of the representation vectors is smallest. 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 pre-segmentation of the relevant object to reduce complexity.

Carbonetto et al. [4] 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 image representations by means of segments. But with our model-free solution we can cope with more complex images.

III. 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 pre-segmented, the location in the image and the viewpoint 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. Having such a classifier, the localization of the object in the image should be straightforward. The learning procedure in our framework (see figure 1) works as follows: The labeled images are put through a preprocessing step that transforms them to greyscale. Then two kinds of regions are detected. On one hand regions of discontinuity are extracted. These are quadratic regions around salient points, 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 [6] with Similarity-Measure-Segmentation. This new segmentation method allows the segmentation of non-connected regions and performs equally or better than several other methods with respect to object recognition in our experiments. In the next step, 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 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 inhomogeneous sets of image features. A further advantage of Boosting is that weak classifiers are calculated when needed instead of calculating unnecessary weak hypotheses a priori. The result of the training procedure is saved as the final hypothesis.

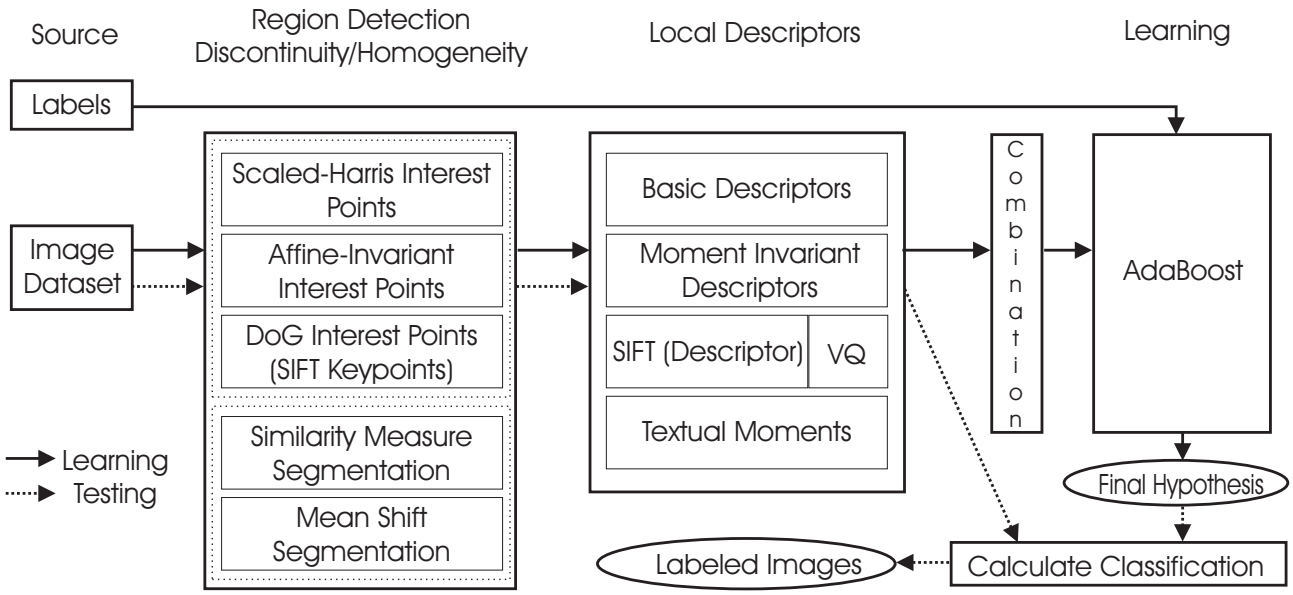


Fig. 1. Our framework for generic object recognition starts from a labeled image database. Regions of discontinuity and homogeneity are extracted and described by local descriptors forming a feature vector. Learning by AdaBoost [12], leads to a final hypothesis which consists of several weak hypotheses.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

Existing datasets for object recognition used by other research groups (e.g. [8], [2]) show the objects with just small variations in scale and objects are generally viewed at similar poses. Figure 2 shows some examples of the Caltech database of the categories cars(rear), motorbikes and airplanes. To aim at a solution for a more generic object recognition we were in need of images having many different instances of the object category at various locations in the image, at different scales, viewed from several positions and with high background clutter. Therefore we have built up our own more complex database. This database[2] (further on termed GRAZ-01) that was used in [28], 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). Figure 3 shows some example images of each category.

Based on our localization results (see section VII-C), 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[2] (further on termed GRAZ-02). This database has been carefully balanced with respect to background, so that similar backgrounds occur for all categories. Furthermore, we increased the complexity of the object appearances and added a third category of images. The database contains 311 images of category person (P), 365 of category bike (B), 420 of category cars (C) and 380 of a counter-class (N, meaning it contains no bikes, no persons and no cars). Figure 4 shows some example images.

To be comparable with existing well known approaches we also carried out experiments on the same

[2]available at http://www.emt.tugraz.at/∼pinz/data/

Fig. 2. Some examples of the Caltech database, categories cars(rear), motorbikes and airplanes, used in [8].

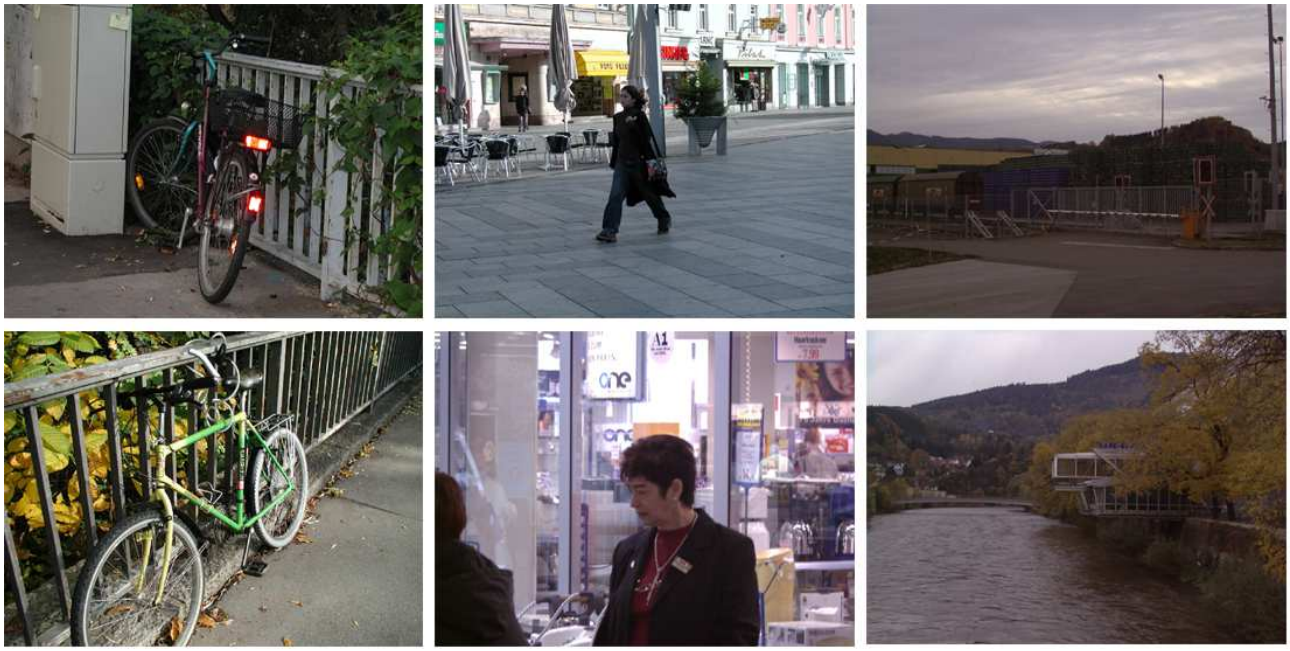


Fig. 3. Some example images from our database GRAZ-01. The first column shows examples of the category bikes (B), in the second column there are images of the category person (P). The rightmost column shows images of the counter-class (N). All these images were correctly classified using our approach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

database like Fergus et al. [8][3] (further on termed Caltech).

Regarding different region detection and description techniques shown in figure 1, we experimentally evaluate two kinds of methods. First we perform various experiments for one region extraction with one kind of local description technique. Not all possibilities are tried out in this work, but we focus on methods with high performance based on results reported in [27] and [28]. The second method is the combination of various kinds of region detections with different description techniques in one learning step (using the "Combination" module shown in figure 1).

The performance is measured by the commonly used receiver-operating-characteristic(ROC)-equal error rate. This error rate is calculated as the percentage of the area under the ROC-curve (for details see [2]).

[3]available at
$http://www.robots.ox.ac.uk/\sim vgg/data/$
and Cars (Side) from
$http://l2r.cs.uiuc.edu/\sim cogcomp/index\_research.html$.

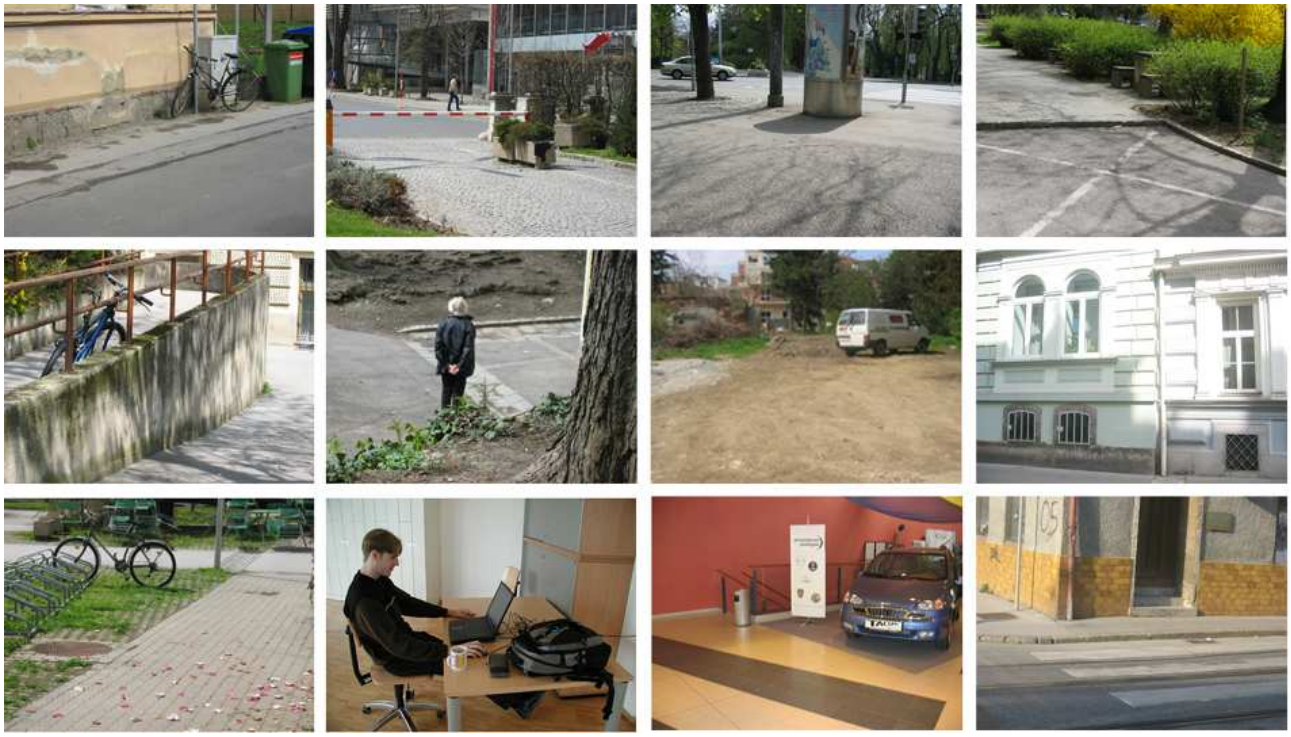


Fig. 4. Some example images from our database GRAZ-02. The first column shows examples of the category bikes (B), in the second column there are images of the category person (P) followed by images of the category cars (C) in the third column. The rightmost column shows some images of the counter-class (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 VII).

IV. 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 category 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], [2], [34]) use just one kind of description method, we combine multiple information extraction methods to capture the essential characteristics of various object categories (e.g. persons, cars, etc). 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.

A. 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], [33], [38]), and at varying scales (e.g. [19], [23], [24]). Based on the evaluation of interest point detectors by Schmid et al. [29], we decided to use the scale invariant Harris-Laplace detector [23] and the affine invariant interest point detector [24], both by

Mikolajczyk and Schmid. In addition we use Lowe's DoG (difference of Gaussian) keypoint detector [20] which is strongly related to SIFTs as local descriptors.

The scale invariant Harris-Laplace detector finds interest points by calculating a scaled version of the second moment matrix M. It localizes points where the Harris-Measure $H = det(M) - \alpha trace^2(M)$ is above a certain threshold th. The characteristic scale for each of these points is obtained by the construction of a Laplacian scale-space. The affine invariant interest point detector uses again the second moment matrix and a Laplacian scale-space. Additionally, an iterative algorithm is used which converges to affine invariant points by modifying the location, scale and neighborhood of each point [24].

To normalize the regions around these two kinds of interest points we have to consider illumination, scale and, in case of the affine invariant detector, affine transformations. For the size normalization, we have decided to use quadratic patches with a side length of l pixels. We extract a window of size $w = 6 \cdot \sigma_I$ (ajar to the value used by Mikolajczyk and Schmid in [24]) where σ_I is the characteristic scale of the interest point delivered by the interest point detector. Scale normalization is achieved by smoothing and subsampling in cases of $l < w$ and by linear interpolation otherwise. In order to obtain affine invariant patches the values of the transformation matrix resulting from the affine invariant interest point detector are used to normalize the window to the shape of a square before size normalization.

For illumination normalization we use Homomorphic Filtering (see e.g. [14], chapter 4.5). The Homomorphic Filter is based on an image formation model where the image intensity $I(x, y) = i(x, y)r(x, y)$ is modeled as the product of illumination $i(x, y)$ and reflectance $r(x, y)$. Elimination of the illumination part leads to a normalization. This is achieved by applying a Fast Fourier Transform to the logarithm image $ln(I)$. Skipping the low Fourier coefficients leads to a separation of the reflectance component (high-pass filer). After the inverse transformation and an exponentiation we get the desired normalized patch.

Lowe introduced an interest point detector which is invariant to translation, scaling and rotation and is minimally affected by small distortions and noise [20]. We use the binary from D. Lowe that already exports the local descriptors of a circular region with a radius of 8 pixels (and 8 orientation planes) around the detected interest points.

B. 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. [5] and [32]) are trying to split images into background and prominent foreground objects. Barnard et al. [3] use these segmentation methods for learning 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 V).

We have developed a new algorithm - "Similarity-Measure-Segmentation" (first presented in [13]) - which is described in detail subsequently. We compare its performance for object categorization with the well known Mean-Shift algorithm by Comaniciu and Meer [6]. In our framework we use the code from "The Robust Image Understanding Laboratory"[4].

1) Similarity-Measure-Segmentation: Similar to other segmentation algorithms (see [5] and [32]), 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 (see equation 1) to describe pixel similarity for segmentation purpose.

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

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

This similarity is used to split images into regions. SC_i defines an element of the Similarity-Criteria vector SC, in other words the distance of two pixels corresponding to a defined pixel feature. The parameters a_i can be set between 0 and 1 to change the weight of the Similarity-Criterion. σ_i is used to change the sensitivity of the different Similarity-Criteria. For example on images with a small intensity variation, a small σ_i is used to enhance the sensitivity of the intensity Similarity-Criterion.

We are extracting two kinds of features. On one hand color, intensity, brightness and the position of a pixel, which consider only a single pixel. And on the other hand Local Binary Patterns (see [26]), high-pass and Wavelets (see [1]), which consider a certain neighborhood of a pixel. The first group of features are easy and fast to calculate but can only be used for regions of homogeneity without texture. The second group is much more complex to calculate, especially Wavelets, but delivers certain texture information.

Equation 2 shows the definition for one similarity criterion element SC_i, where $P1_i$ and $P2_i$ are the two pixels and i is the index of the used feature. We use the Euclidean distance to calculate the elements of SC.

$$SC_i = \|P1_i - P2_i\| \tag{2}$$

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

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 the Mean-Shift segmentation [6]).

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

$$r_{max} = \ln(\frac{t}{\sum_{i=1}^{n} a_i}(n-1))(-2\pi\sigma_x) \tag{3}$$

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.

Figure 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 algorithms merges the maintaining rail parts to 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, that is split into 11 small regions

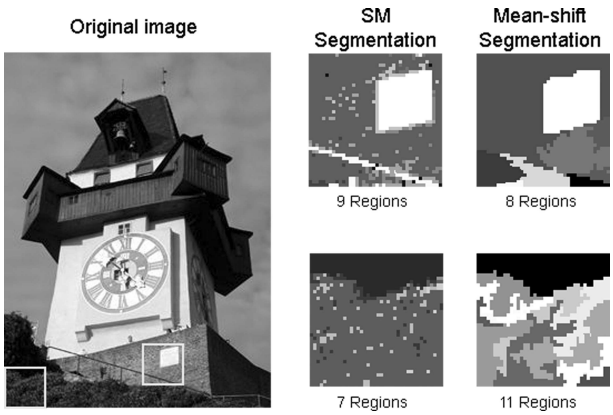


Fig. 5. Two detail views of the "Grazer Clocktower" segmented with Similarity-Measure segmentation (images in the middle) and mean-shift segmentation (images on the right).

with Mean-Shift segmentation, and 5 disconnected regions surrounded by 2 large regions with Similarity-Measure segmentation. The results are desirable for our purpose because it turns out that a representation by not connected regions leads to a better performance of our categorization approach.

V. Local Description

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

For regions of discontinuity, local descriptors have been researched quite well (e.g. [11], [20], [30], [15]). We selected four local descriptors. Our first descriptor is simply a vector of all pixels in a patch subsampled by two. The dimension of this vector is $\frac{l}{4}^2$, which is rather high and increases computational complexity. As a second descriptor we use intensity moments $M^a_{I_{pq}} = \int\int_\omega i(x,y)^a x^p y^q \, dx \, dy$ 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. In view of the performance evaluation of local descriptors done by Mikolajczyk and Schmid [25] we took SIFTs (see [20]) as a third and Moment Invariants (see [15]) as a fourth choice. In [25] the SIFTs outperformed the other descriptors in nearly all tests and the Moment Invariants were average for all considered aspects.

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.

As shown in [20] 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.

We use two groups of descriptors for the representation of regions of homogeneity. The first group describes the intensity or the color 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 group contains invariant moments (see [22]), 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 I gives an overview of the various description methods in our framework and their dimension.

VI. Learning Model

Our learning model is based on the AdaBoost algorithm [12]. This algorithm was adapted by adding the possibility of putting different weights on positive and negative training images. We set up a new

TABLE I

THIS TABLE GIVES AN OVERVIEW OF THE DESCRIPTION METHODS IN OUR FRAMEWORK AND THEIR DIMENSION (FOR REGION SIZE OF $16 \times 16 pixels$).

-	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 $(I_1, \ell_1), \ldots, (I_m, \ell_m)$.
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 $w_1, \ldots, w_m$ 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 $\beta_t = \sqrt{\frac{1-\varepsilon}{\varepsilon}}$.
4) Update $w_k \leftarrow w_k \cdot \beta^{-\ell_k \cdot h_t(I_k)}$ for $k = 1, \ldots, m$.

Output the final hypothesis (classifier):

$$H(I) = \text{sign}\left(\sum_{t=1}^{T} (\ln \beta_t) h_t(I)\right)$$

Fig. 6. Shows the standard AdaBoost algorithm [12].

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 want 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 $(I_1, \ell_1), \ldots, (I_m, \ell_m)$ 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 (see figure 6).

A. 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.

$$\sum_{k=1, h(I_k)=\ell_k}^{m} w_k > \sum_{k=1, h(I_k)^m \neq \ell_k} w_k \tag{4}$$

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 a very weak requirement. The process of putting weights and constructing a weak hypothesis is iterated for several rounds $t = 1, \ldots, T$, and the weak hypotheses h_t of each round are combined into the final hypothesis H (for details see figure 6).

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 equation (4) — 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}[5], 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.

Obviously AdaBoost is a very general learning technique for obtaining 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 $\mathcal{R}(I) = \{(\tau_i, v)\}$ where τ_i 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.

B. Weak-Hypotheses-Finder

Using one type of description vector at a time is the basic functionality of our learning algorithm. Using various description methods τ_i separately simplifies our learning algorithm to a case where we represent an image I_k by a set of descriptors $\mathcal{R}(I_k) = \{v_f\}$, $f = 1, \ldots, F_k$, where v_f denotes the values of the descriptor as real vector, and F_k is the number of extracted description vectors in an image I_k (see figure 7 for an explanation of this Weak-Hypotheses-Finder using just one description method). The weak hypotheses for AdaBoost are calculated from these descriptors. 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 value v and a similarity threshold θ. The threshold θ decides if an image contains a description vector v_f that is sufficiently similar to v. The similarity between v_f and v is calculated by the Mahalanobis distance for Moment Invariants, basic intensity moments and the descriptors for the regions of homogeneity. This distance is chosen because it is measured in terms of standard deviations from the mean of the training samples. For SIFTs and the subsampled grayvalues the Euclidean Distance is used. This simpler measure is sufficient in this case because of the bounded domain of the description vector values between 0 and 255. The Weak-Hypotheses-Finder (figure 7(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 (see figure 7(5)).

The main computational burden is the calculation of the distances between v_f and v (see figure 7(2)), since they both range over all description vectors that appear in the training images. We arrange the

[5]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 $(\mathcal{R}(I_k), \ell_k)$, $k = 1, \ldots, m$, $\mathcal{R}(I_k) = \{(v_{k,f} | f = 1, \ldots, F_k)(w_{k_0})\}$.
(1): Distance functions: Let $d(\cdot, \cdot)$ be the distance in respect to the description vectors in the training images.
(2): Minimal distance matrix: For all description vectors $v_{k,f}$ and all images I_j calculate the minimal distance between $v_{k,f}$ and description vectors in I_j,

$$d_{k,f,j} = \min_{1 \leq g \leq F_j} d(v_{k,f}, v_{j,g}) \ .$$

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

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

(4): Select best weak hypothesis (Scanline): For all description vectors $v_{k,f}$ 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 $v_{k,f}$ 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

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

Fig. 7. Explanation of the Weak-Hypotheses-Finder using just one description method.

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 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 dataset of 150 positive and 150 negative images. Each image has an average number of approximately 400 description vectors. After preprocessing, using SIFTs one iteration of Boosting requires about ten seconds computation time on a P4 (2.4GHz PC).

C. Weak-Hypotheses-Finder with Multiple Description Methods

In an outlook in [28] we discussed the possibility of using multiple description methods in one learning procedure. The challenge of the learning algorithm is now 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. Using these multiple description methods at a time changes the algorithm in the following way. An image I_k is now represented by a list of description methods $(\tau_{k,l}, v_{k,f})$, $f = 1, \ldots, F_k \wedge l = 1, \ldots, n$, where $\tau_{k,l}$ denotes the type of a feature, n denotes the number of the different used description methods, $v_{k,f}$ denotes again its value as real vector, and F_k is the number of extracted descriptors in an image. Therefore, a weak hypothesis h has to select a feature type τ and again its value v, and a similarity threshold θ. Details how the Weak-Hypotheses-Finder changes when using multiple description methods are given in figure 8.

Obviously the computational complexity is increasing with every additional kind of feature used.

VII. Experiments and Results

The experimental evaluation is split into three parts. The first part (subsection VII-A) specifies the parameter settings. Our classification results are discussed in detail in subsection VII-B, showing that this

Input: Labeled representations $(\mathcal{R}(I_k), \ell_k)$, $k = 1, \ldots, m$, $\mathcal{R}(I_k) \triangleq \{(\tau_{k,f}, v_{k,f}) | f = 1, \ldots, F_k, (w_{k_0})\}$.
(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 maxium sum the threshold θ is set to

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

Fig. 8. Explanation of the Weak-Hypotheses-Finder using various description methods at a time.

approach clearly outperforms current state-of-the-art techniques. The benefits of using multiple features in one learning procedure are also pointed out there. Subsection VII-C 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.

A. 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 $th = 30000$ 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 [27]). 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 [20]. Initially we took all points into account but then we clustered the SIFT description vectors (8x8 pixels with 8 orientation planes) 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 $cl = 300$[6]) 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

[6]this numbers were experimentally evaluated and depend on the image complexity, for details see **??**

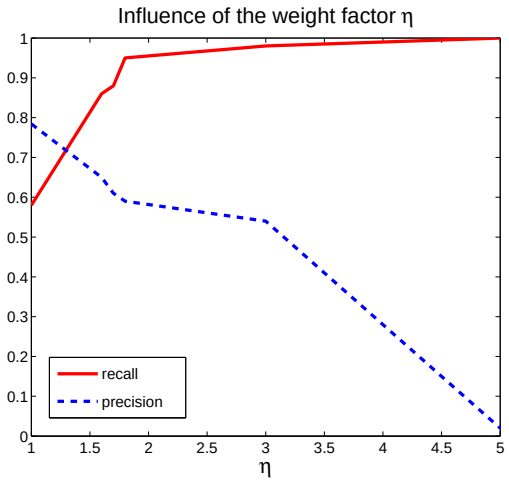


Fig. 9. The diagram shows the influence of an additional factor η for the weights of incorrectly positive classified examples. The recall increases faster than the precision drops until a factor of 1.8 (for the GRAZ-01 dataset with affine invariant regions and Moment Invariants). The optimal value of this factor varies on different datasets.

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 the following parameters: $T = 100$, $\eta = 1.0$. The ROC curves can either be obtained by varying the additional weight factor η or by testing the classifier varying the threshold for classification (both result in the same ROC-equal-error rate).

Figure 9 shows the influence of the additional weight factor on recall and precision. In this test on the bike category of the GRAZ-01 dataset, 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 dataset. For an evaluation with ROC-equal-error rate this factor does not influence the result.

The power of our new Similarity-Measure-Segmentation with respect to object categorization is shown in table II. 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.

TABLE II

RELATIVE ERROR ON DATASET CARS(REAR) (CALTECH DATABASE) AND BIKES (OF GRAZ-01 AND GRAZ-02). 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 $reg_{min} = 250$, THE OBJECT CATEGORIZATION WORKS BETTER WITH SIMILARITY-MEASURE-SEGMENTATION.

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

TABLE III

SHOWS THE ROC-EQUAL ERROR RATES ON THE CALTECH DATABASE AND ON CARS SIDE FROM THE UNIVERSITY OF ILLINOIS. 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.

Dataset	(1)	(2)	[8]	Others
Motorbikes	94.3	92.2	92.5	94.0 [**?**], 88 [37], 93.2 [34]
Airplanes	97.5	88.9	90.2	-
Faces	100	93.5	96.4	93.5 [37]
Cars(side)	100	83.0	88.5	97.5 [?], 79 [2]
Cars(rear)	100	91.1	90.3	93.9 [?], 86.5 [37]

B. Classification Results

1) Reference Dataset: To be comparable with existing approaches we first evaluated our method on the Caltech database. 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[7]. The results are shown in the first column of table III. 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. The comparison with the results of Fergus et al. in the last row of table III shows that our best results are superior to the classification performances of Fergus et al. for all categories, even if we train with fewer images. Especially Similarity-Measure-Segmentation based region detection yields a very significant improvement compared to the results of Fergus et al.

2) GRAZ-01 dataset: Having demonstrated the good performance of our approach on a reference dataset (Caltech), we proceeded with experiments on our first database with higher complexity, the 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). 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 IV 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. Figure 10 shows the recall-precision curves of these experiments.

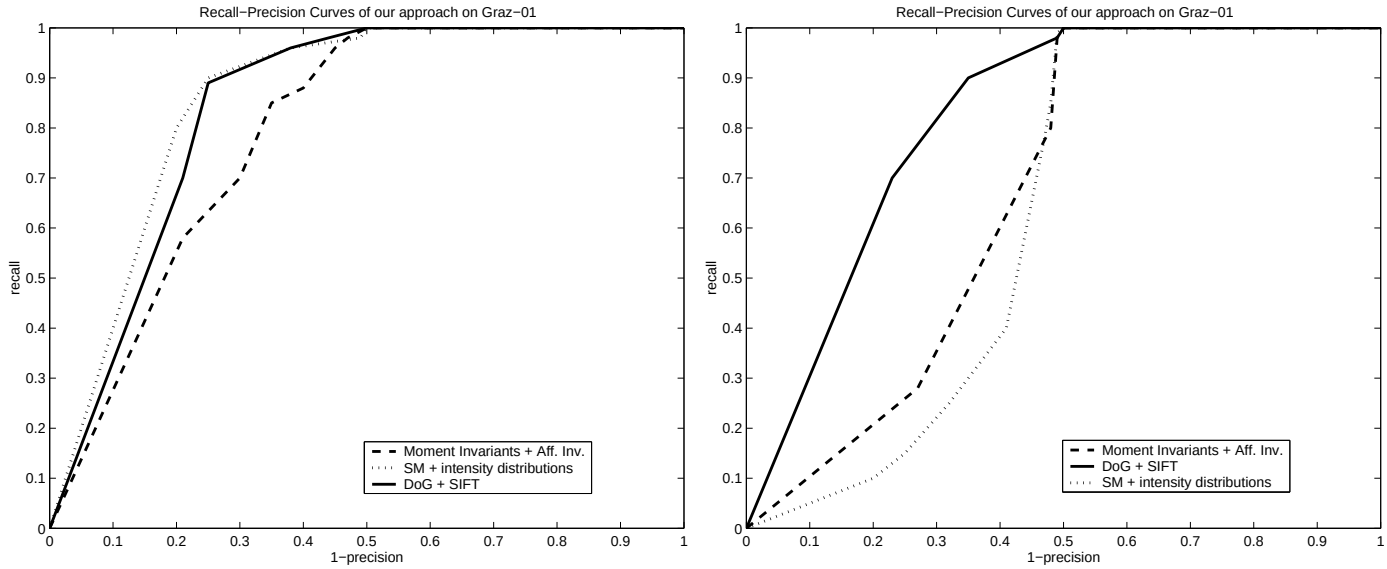
All images presented previously in figure 3 were categorized correctly. Figure 11 gives examples of incorrectly classified images. In both cases the images of the counter-class result from an experiment

[7]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.

TABLE IV

THIS TABLE SHOWS A COMPARISON OF THE ROC-EQUAL ERROR RATES 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
Bikes	76.5	86.5	89.6
Persons	68.7	80.8	59.1



(a) (b)

Fig. 10. 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.

where we trained the category bikes. For the classification of these example images we used an average classifier with $\eta = 1.0$ and a classification threshold of 0.0 for the testing.

3) GRAZ-02 dataset: After these experiments on the GRAZ-01 dataset we evaluated our approach using the GRAZ-02 dataset. 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. Table V shows the classification performance of various specific combinations of region extractions and description types. 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.

Again, all the images in figure 4 were categorized correctly while images in figure 12 represent examples, where the classification fails. These images where classified with average parameters of $\eta = 1.0$ and a classification threshold of 0.0 for the testing.

A qualitative visual comparison of figures 3 and 11 with figures 4 and 12 immediately reveals the need of further explanation. Although the overall categorization results 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 dataset 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 try to answer some of these questions in subsection VII-C in the light of localization abilities of the various detectors.

Fig. 11. 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 image not containing a bike, in the second column are images of the category person (P) classified as images not containing a person. The rightmost column shows images of the counter-class-set (N) that were classified as bikes (B).

TABLE V

ROC-EQUAL-ERROR RATES OF VARIOUS SPECIFIC COMBINATIONS OF REGION EXTRACTIONS AND DESCRIPTION METHODS ON THE THREE CATEGORIES OF THE GRAZ-02 DATASET. 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.

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

4) Combination: Subsequently, we show experiments performed by using more than one type of the various region extractions with a description method in one learning step. We did three kinds of combinations[8]. The first part are the regions obtained with the affine invariant interest point detection, described with Moment Invariants. We combine it with regions achieved through DoG keypoints described by SIFTs (see table VI (a)), regions extracted with the affine invariant interest point detector described with basic intensity moments (see table VI (b)) and regions of homogeneity obtained by the Similarity-Measure-Segmentation and described with intensity distributions (see table VI (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 74.81 vs. 81.2 for persons). The main benefit is that a 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.

[8]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 rightmost column shows some images of the counter-class-set (N). These are examples that were classified as bikes (B).

TABLE VI

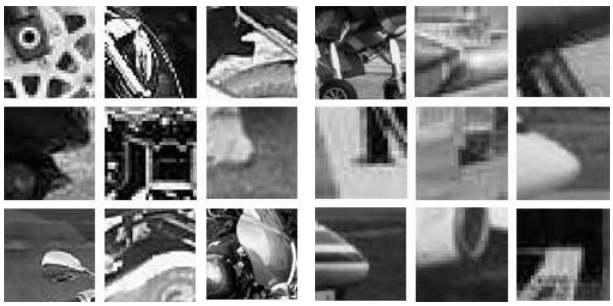
THIS TABLE SHOWS THE ROC-EQUAL ERROR RATES OF SPECIFIC COMBINATIONS OF REGION EXTRACTIONS AND DESCRIPTION METHODS SEPARATED AND THEIR PERFORMANCE IF THEY ARE COMBINED IN ONE LEARNING STEP (ON GRAZ-02). THE FIRST VALUE IS ALWAYS FOR THE MOMENT INVARIANTS. THE SECOND COLUMN SHOWS THE RESULTS OF EITHER 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.

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

C. Localization Performance

To discuss the localization of the information learned by our approach, we first evaluated the experiments shown in the previous subsection 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 III, we found out, that only 25% to 50% 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. Figure 13 shows some examples of regions of homogeneity selected as weak hypotheses from the Caltech dataset. 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 dataset 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

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.



(a) (b)

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.

hand, object localization is not straight forward if we use regions of homogeneity on images with specific background appearances.

Figure 14(a) 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, figure 14(b) shows weak hypotheses of the final classifier of the category airplanes. With this specific combination we obtain 80% to 90% 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 III is lower compared to using regions of homogeneity.

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 IV based on the GRAZ-01 dataset. We observe an average of 60% 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% 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%.

Table VII 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 V 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. Figure 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 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

TABLE VII

THIS TABLE SHOWS THE PERCENTAGE OF THE WEAK HYPOTHESES THAT ARE NOT LOCATED ON THE OBJECT. HERE WE USED THE SAME COMBINATIONS AS IN TABLE V FOR THE GRAZ-02 DATASET.

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

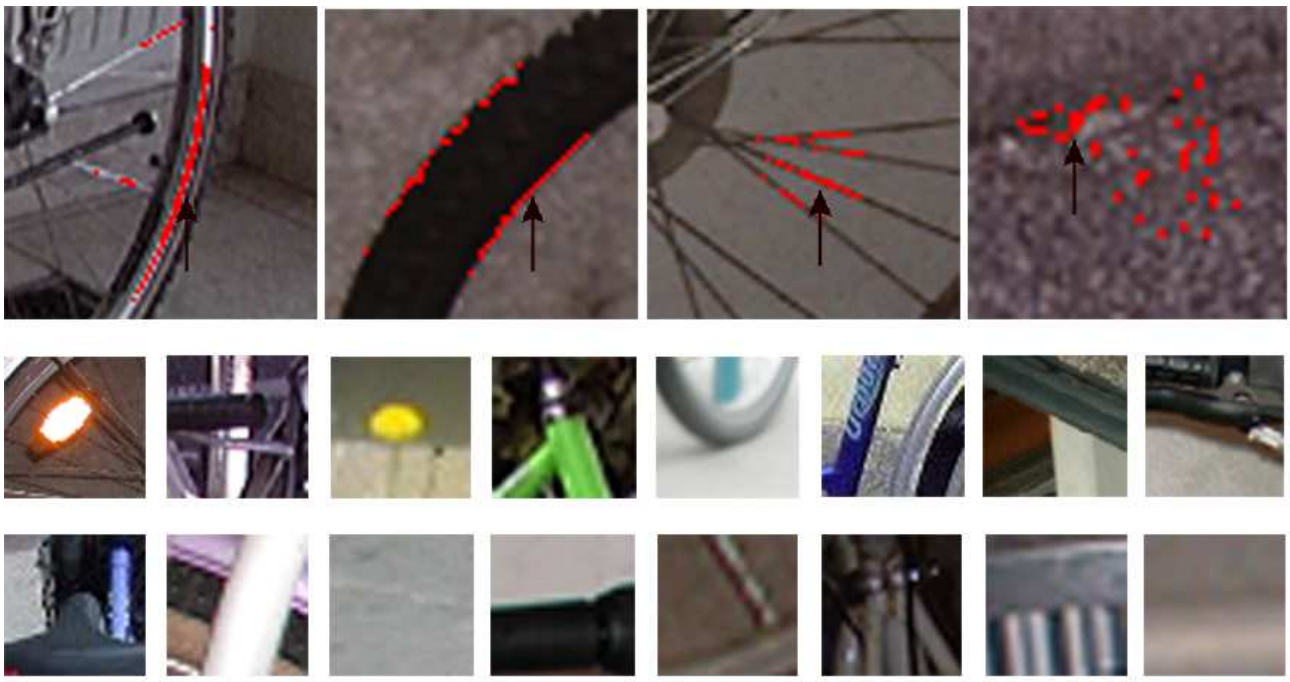


Fig. 15. Shows examples of weak hypotheses used for the final classifier of the category bike (GRAZ-02). The first row shows hypotheses based on the test with regions of homogeneity and intensity distributions. The middle row shows regions extracted with the affine invariant interest point detector and described by Moment Invariants. Examples of weak hypotheses obtained at 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.

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. 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).

VIII. 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. 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. The main contributions of the paper, however, lie in the new concept of learning, the possibility of combining various types of image features and the presentation

of a new segmentation technique (Similarity-Measure-Segmentation). 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. To reduce the complexity of our approach we are considering a reduction of the number of description vectors with a solution that is better fittting than simple k-means clustering. Also the localization problem will be investigated in more detail. The different performances of various combinations in this framework leads to the need of a loop within the learning procedure. There a first unsupervised localization step is followed by the actual learning procedure. 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. 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 of discriminative spatial relations.

ACKNOWLEDGMENT

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

REFERENCES

[1] J. C. Feaveau A. Cohen, I. Daubechies. Biorthogonal bases of compactly supported wavlets. In *Commun. Pure Appl. Math.*, pages 485–560, 1992.

[2] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *Proc. European Conference on Computer Vision*, pages 113–130, 2002.

[3] 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. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 675–682, 2003.

[4] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *Proc. European Conference on Computer Vision*, pages 350 – 362, 2004.

[5] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, volume 24(8), pages 1026–1038, 2002.

[6] D. Comaniciu and P. Meer. Mean shift: A robust approach towards feature space analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 24(5), pages 603–619, 2002.

[7] Gy. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *Proc. International Conference on Computer Vision*, pages 634–640, 2003.

[8] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2003.

[9] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *Proc. European Conference of Computer Vision*, pages 242–256, 2004.

[10] V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneos object recognition and segmentation by image exploration. In *Proc. European Conference on Computer Vision*, pages 40–54, 2004.

[11] W. Freeman and E. Adelson. The design and use of steerable filters. In *PAMI*, pages 891–906, 1991.

[12] Y. Freund and R. Schapire. A decision theoretic generalisation of online learning. *Computer and System Sciences*, 55(1):119–139, 1997.

[13] M. Fussenegger, A. Opelt, A. Pinz, and P. Auer. Object recognition using segmentation for feature detection. In *International Conference on 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. In *Proc. European Conference on Computer Vision*, pages 642 – 651, 1996.
[16] R. M. Haralick. Statistical and structural approaches to texture. In *Proc. Conference of Pattern Recognition*, volume 67, pages 786–804, 1979.
[17] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of the* 4^{th} *Alvey Vision Conference*, pages 189–192, Manchester, 1988.
[18] R. Laganiere. A morphological operator for corner detection. *Pattern Recognition*, 31(11):1643 – 1652, 1998.
[19] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.
[20] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, pages 1150–1157, 1999. The binary is available at http://www.cs.ubc.ca/ lowe/keypoints/.
[21] W. Maass and M. Warmuth. Efficient learning with virtual threshold gates. *Information and Computation*, 141(1):66–83, 1998.
[22] S. Maitra. Moment invariants. In *Proc. IEEE 67*, pages 679–699, 1979.
[23] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. International Conference on Computer Vision*, pages 525–531, 2001. The binary is available at: http://lear.inrialpes.fr/people/Mikolajczyk/.
[24] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. European Conference on Computer Vision*, pages 128–142, 2002. The binary is available at: http://lear.inrialpes.fr/people/Mikolajczyk/.
[25] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proc. Conference on Computer Vision and Pattern Recognition*, volume 2, pages 257–263, 2003.
[26] T. Ojala and M. Pietikinen. Unsupervised texture segmentation using feature distributions. In *Journal of Pattern Recognition*, volume 32, pages 477–486, 1999.
[27] A. Opelt. Feature selection for scaled interest points. Master's thesis, Graz University of Technology, 2003.
[28] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proc. European Conference on Computer Vision*, volume 2, pages 71–84, 2004.
[29] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, pages 151–172, 2000.
[30] Cordelia Schmid and Roger Mohr. Local greyvalue invariants for image retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.
[31] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *International Journal of Computer Vision*, 56(3):151–177, 2004.
[32] J. Shi and J. Malik. Normalized cuts and image segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 22(8), pages 888–905, 2000.
[33] E. Shilat, M. Werman, and Y. Gdalyahu. Ridge's corner detection and correspondence. In *Proc. Computer Vision and Pattern Recognition*, pages 976 – 981, 1997.
[34] J. Thureson and S. Carlsson. Appearance based qualitative image description for object class recognition. In *Proc. European Conference of Computer Vision*, pages 518–529, 2004.
[35] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 511–518, 2001.
[36] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *Proc. International Conference on Computer Vision*, pages 257–264, 2003.
[37] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. European Conference on Computer Vision*, 2000.
[38] R. P. Wuertz and T. Lourens. Corner detection in color images by multiscale combination of end-stopped cortical cells. In *International Conference on Artificial Neuronal Networks*, pages 901 – 906, 1997.