ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG
INSTITUT FÜR INFORMATIK

Lehrstuhl für Mustererkennung und Bildverarbeitung

Patch Based Approaches for the Recognition of Visual Object Classes - A Survey

Internal Report 2/06

Alexandra Teynor
November, 2006

Patch Based Approaches for Visual Object Class Recognition - a Survey

Alexandra Teynor

November 29, 2006

1 Introduction

Today, a huge amount of digital images exists on private computers, the web, in digital galleries or professional media archives, and the mass will continue to grow. Current image databases of publishing companies already reach many thousands of terrabytes.

Retrieving images again from these databases is difficult. Think about your own holiday picture collection: manual search through archives is hardly feasible on home desktop computers (if at all), not to speak of professional archives with terrabytes of image data. Manual annotation of images and then using text retrieval techniques on the key words is no general solution either. First of all, it is very labor and with that cost intensive, and so only feasible for companies. It is subjective, since visual content tends to be interpreted in different ways by humans. Other drawbacks are the language dependency of the key words, the possibility of spelling mistakes or simply the use of wrong key words. Professional image classification schemes as used for physical image archives like the ICONCLASS scheme try to overcome some of these difficulties by formalizing the image description, however, they are difficult to handle by non experts.

We need automatic techniques to cope with the sheer amount of data. Of course, there are different tasks to be solved. One of the first challenges met in this context was the search for “similar images”, where the notion of “similar” was mainly defined from a color, texture and sometimes shape point of view. The images were treated more or less globally, later also local considerations came into play. This problem was heavily researched in the field of “Content Based Image Retrieval” (CBIR) in the 1990s and early 2000s, and work still continues on this topic. In a CBIR system, typically one or more example images are used as input, for which then images are retrieved sorted according to their relevance. This search method is known as “query-by-example” paradigm, but there exist also others, like “query-by-sketch” or “query-by-color” to just name a few.

However, a more common problem is the search for specific objects. Users are more interested in finding semantic entities in images like people, cars or animals. Here we can distinguish between two main cases: on one hand the detection of the very same physical object in an image, on the other hand the recognition of members of object classes. Of course, object classes can be broad or narrow, and in order to be recognized by their appearance, they have to share some visual characteristics. So the common notion for this query type is to search for “visual object classes”.

Eakins [17] proposed a classification scheme of query types consisting of three levels of increasing complexity. In this scheme, the search for visual object classes falls into level 2. The partition of the query types were made as follows:

1. **Level 1: Search by primitive features**
   Images are retrieved by basic features like color, texture, shape, spatial layout or combination of these. Most traditional image retrieval systems, as e.g. QBIC [21], SIMBA [60], VIPER/GIFT [48], or FIRE [11] work on this level. All information necessary can be aquired from the images themselves.

2. **Level 2: Search by derived/logical features**
   Additional knowledge is necessary for the retrieval of correct images, e.g. that a certain structure has been given a specific name or that a visual object class has certain properties. The subtasks falling into this category are:

   a) *Retrieval of objects of a given type*
      In this category, visual object classes are searched for. In current research, the object classes are rather narrow, however also more general classes like, e.g., “flowers” or “animals” could be considered.

   b) *Retrieval of individual objects or persons*
      Here, exactly the same instance of an object or person should be retrieved. Even if this task sounds more difficult, since, e.g., not only any car, but a special car is searched for, this is an easier task for an computer, since less variety in appearance has to be taken into account.

3. **Level 3: Search by abstract features**
   Here, the meaning and purpose of images should be judged, so high level reasoning is necessary. To our knowledge, currently no systems are really working at this level, besides maybe interpreting the meaning of the prevalent color in an image [9]. Again, two subtypes can be distinguished:

   a) *Retrieval of named events or types of activity*
      The visual variety of images associated to a football match or an event like, e.g., the “Oktoberfest” is enormous, so learning is difficult in this area.

   b) *Retrieval of pictures with emotional or religious significance*
      Here, the mood and/or meaning of images should be considered, something that even humans do not easily agree on. This stage of image retrieval is not likely to be solved by machines in the near future, if at all, since it requires some higher degree of intelligence.

2 Terminology

First of all, we have to clarify the meaning of certain terms used throughout this work, since they are often used differently in literature. It is important to state our interpretation of these terms:

Object: *“something material that may be perceived by the senses” [1]*
In our work, we only deal with physical objects that can be counted. E.g., we would not consider “snow” an object, but a “snow crystal”. The objects should also have limited extents or they have to be sufficiently far away, so that they can be captured by images.

(Visual) Object class: *"a collection of objects that share some visual characteristics"*
Examples for a visual class would be "cars", "faces" or "motorbikes", since they all have certain visual properties in common, despite their sometimes large variance in appearance. The object class "tools" would not fall into this category, since tools might look completely different, and the grouping is made form a function point of view.

Classification: *"systematic arrangement in groups or categories according to established criteria"[1]*
In our case, images are assigned to a predefined number of classes. There might be as little as two classes, e.g. in a simple object present/absent task, or many thousands, e.g. for automatic image annotation, where key words from a large pool have to be assigned to the images. In order to cope with previously unseen classes, an additional class holding the "unknown" entities might be established. One image might be assigned to different classes, in case different objects can be found, or different key words apply. For most experiments however, only one object class is to be recognized per image. An alternative term used in literature is "categorization"[31].

Identification: *"idetification is recognition of an individual object within an class"[69]*
Identification means not only to recognize any member of a class, e.g. a face, but a specific instance of that class, say my face. The object class involved is determined implicitly with that.

Recognition: *"the act to perceive to be something or someone previously known" [1]*
The term "recognition" is used in two contexts: on one hand for specific objects, then this term is equal to "identification". On the other hand for generic objects [53] or object classes, then "classification" is meant. We will explicitly state which notion of recognition we are using if not immediately clear form the context.

Localization: *"the discovery of the exact position of a given object in an image"*
Typically, the location and the extent, sometimes also the orientation of the object has to be detected. In a simple case, it is previously known (or assumed) that an instance of the object class is present, and the most probable position is calculated. In other cases, recognition and localization are coupled: the presence of a certain object class is determined by deciding whether an object at a certain location, scale and or orientation can be found. In the simplest case, a sliding window is applied to the image and the subwindows are classified. As a result, a bounding box can be drawn around probable object locations, other approaches even deliver segmented objects [31]. "Detection" is sometimes used interchangeably for localization (e.g. [69]).

Detection: This term is used in a lot of context and thus the most ambiguous in this list. It is used in the same was as "recognition", both for visual classes [10] as well as specific objects. On the other hand, it is sometimes synonymous for "localization" [69].

Appearance based methods: *"features used for classification should be extracted from the visual appearance of the image/object in question"*
Keysers [28] states that appearance based methods only use pixel intensities themselves, which can possibly be preprocessed (e.g. brightness corrected). However, most researchers [60] consider any features calculated from the original image as "appearance based", even if more complex functions are involved and we also hold this opinion. Even Keysers emphasizes that the term "appearance based" should mainly establish a border to segmentation based approaches, not to general feature extraction techniques.

Figure 1: A specific object (here a VW beetle) from different views (from [5])

3 Problem Statement

In this work, we deal with classification and localization of visual object class members. In most cases, the task to be solved is to determine whether an instance of an object class is present in an image or not. This question sounds easy, since for humans this is a very easy task, and children at the age of 2 years are already able to recognize many object categories. However, for a computer, this is a very hard problem. Let us illustrate where the difficulties in this very general question lie.

3.1 Difficulty

Why is the recognition of visual object class members in images so difficult for a computer?

- **3D objects in 2D images**
  In the real world, we deal with 3D objects. When they are projected onto a 2D image, information is lost necessarily, since not all views of the object can be captured at the same time. A car looks very different seen from the front, rear, the side or from above, as can be seen in Figure 1. We humans have no difficulty in recognizing these objects even so, since we know all views and how they are related. We have a 3D model of the object class in mind. A general object recognition system would also require this information. One could either supply it with a full 3D model (e.g. [54]), or with a sufficient number of training samples showing different views. To our knowledge, 3D models have only be constructed for the recognition of specific objects, not for object classes so far. A more restrictive but widely adopted approach is to limit the search to a specific view, e.g. exclusively side views of cars or frontal views of faces [18].
- **Projections and geometric transformations**
  When we photograph an object, we project the 3D item onto a 2D image. Different images from the same object are related by a homography [25]. To make things more tracktable, one often assumes planar objects, or at least planar object parts. If we only consider small patches on the surface of

Figure 2: Object variability for the rear of a car

an object, this is approximately true. For planar objects, we have to cope with translations as well as similarity, Euclidean, affine or projective transformations. To simplify things, we usually assume infinite cameras and with that transformations only up to affinity, since they are more easy to handle mathematically.

Let $\mathbf{x} = (x_x, x_y)$, $x_x, x_y \in \mathbb{R}$ be the coordinates of a pixel in an image, then the new coordinates $\mathbf{x}'$ of the transformed point are:

$$\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{t} \tag{1}$$

where the type of the transformation depends on the properties of $\mathbf{A}$:

$\mathbf{A}$	=	$\mathbf{I}$	**translation**
$\mathbf{A}^T\mathbf{A}$	=	$k\mathbf{I}$	**similarity transformation**
$\mathbf{A}^T\mathbf{A}$	=	$\mathbf{I}$	and
$\mathbf{A}^{-1}$	=	$\mathbf{A}^T$	**Euclidean transformation**
$\det(\mathbf{A})$	$\neq$	0	**affine transformation**

These geometric distortions make a direct comparison of images of even the same object difficult, as can again be seen from image pairs in Figure 1.

- **Occlusions**
  In many real world photographs, the object is only partially visible, since it is occluded to some extent, or some parts of the object stretch beyond the image border.

- **Intra class variability**
  The objects typically have a great variability in appearance and layout of the parts. Even if we have a very narrow object category, e.g. “car”, and they are all viewed from the same perspective, e.g., the rear, they can look rather different in detail (see Figure 2).

- **Non rigid transformations of the object itself**
  Some objects are composed of articulated parts, e.g., humans, which makes recognition according to the shape difficult. Other objects have a “soft” structure with no specific outline, e.g., clouds or toy animals.

- **Recording procedure**
  The recording process also introduces errors. These can, e.g., be noise, quantization errors, discretization errors, image blur, but also compression artefacts.

- **Illumination changes**
  Objects captured in the real world might be illuminated very differently. We have to deal with additive (the basic brightness is higher), multiplicative (higher contrast) and non-linear (light source at a different direction) illumination changes.

Figure 3: Images form the semantic class VW beetle (from [5])

- **Ratio image area/object area**
  The object might only cover a small part of the image, while background clutter or other objects dominate the scene. This makes the recognition of these small objects very difficult, especially if we have no a priori information about the scale of the object.

- **Inadequacies of the mathematical model**
  When modelling is an issue, we usually have to make simplifying assumptions about some conditions, in order to keep the problem manageable. In reality however, we might have different conditions, e.g., non linearities, non planarities or statisical dependencies where we assumed none.

- **Semantic notion**
  Even if we only want to consider objects that share some optical characteristics, the amount of visual resemblance can still vary. As human beings, we always have a semantic interpretation of what we see. This fact is known form CBIR and called the "semantic gap". In [55], (p. 1353) it is defined as:

  "The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation."

  How different images with the same semantic interpretation (here VW beetle) might look can be seen in Figure 3. Not only the views and display details are different, but also the styles in which the pictures are made. Here we rather deal with a semantic than with a visual object class.

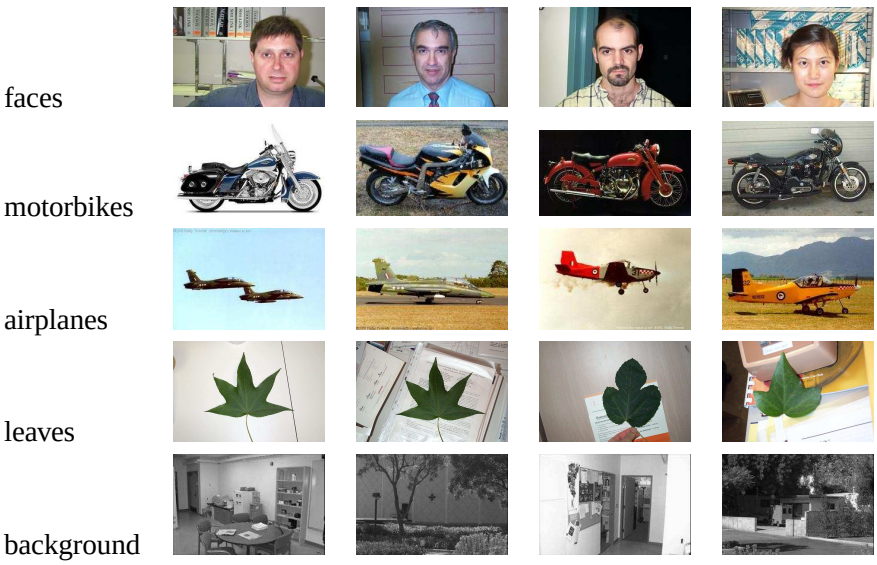
As can be seen from the collection above, a variety of things has to be considered when the recognition of visual object class members should be successful. In current systems, typically only certain aspects are worked on, and the databases currently used make some simplifying assumptions, e.g., about the location, size and/or orientation of the object. This is necessary in order to better control the effects of certain algorithms, but it also restricts the portability of the results. However, more and more reference databases at greater levels of difficulty exist. Some current databases are introduced in Section 3.2.

3.2 Databases

This Section briefly introduces some databases at different levels of difficulty, which are widely used. For these, a lot of reference results are available.

3.2.1 Caltech Datasets

The Californian Institute of Technology (Caltech) offers a set of image collections for object class recognition. They are available on the institutes website[1]. The most commonly used collections are “airplanes_side” (800 images), “faces” (450 images) and “motorbikes_side” (800 images), others are the “cars rear” as well as the “leaves” database, where the latter is comprised of 3 different leave types. For these database, an object present/absent task has to be solved, specific training and test sets are available for better comparability of the results. As a counter class, a set of mixed “background” images is used, except for the cars_rear task, where street backgrounds are provided. The individual objects differ in appearance and location, but are about the same size and orientation. The background is cluttered.



For multi class object recognition, two other datasets are provided, the Caltech 101 and the Caltech 256 object class database. There, images for 101/256 object classes are provided, with 40-800 images per category. The images are all about the size of 300 x 200 pixels.

3.2.2 Graz Datasets

A clearly more difficult categorization task is present in the Graz02 database[2] introduced by Opelt et al. [51, 50]. This database has three object categories: “cars” (420), “persons” (311 images), “bikes” (365 images) and a so-called “none” category (380 images) which is used as a counter class. In all the categories, objects suffer from severe occlusions and have a highly variable appearance and pose, reflecting real world

[1] http://www.robots.ox.ac.uk/~vgg/data3.html
[2] http://www.emt.tugraz.at/~pinz/data/GRAZ_02

scenes more accurately. Here also, an object present/absent task has to be solved.

3.2.3 Eth80 Dataset

The ETH80 database[3] was introduced by Leibe et al. in [30]. Here, 10 different objects from 8 different object classes are photographed in front of a uniform background. For each object, 41 views are taken at different angles. In the following graphic, all individual objects are shown in a refernce position. Tests are usually performed in a leave-one-object-out approach.

[3] http://www.mis.informatik.tu-darmstadt.de/Research/Projects/categorization/download.html

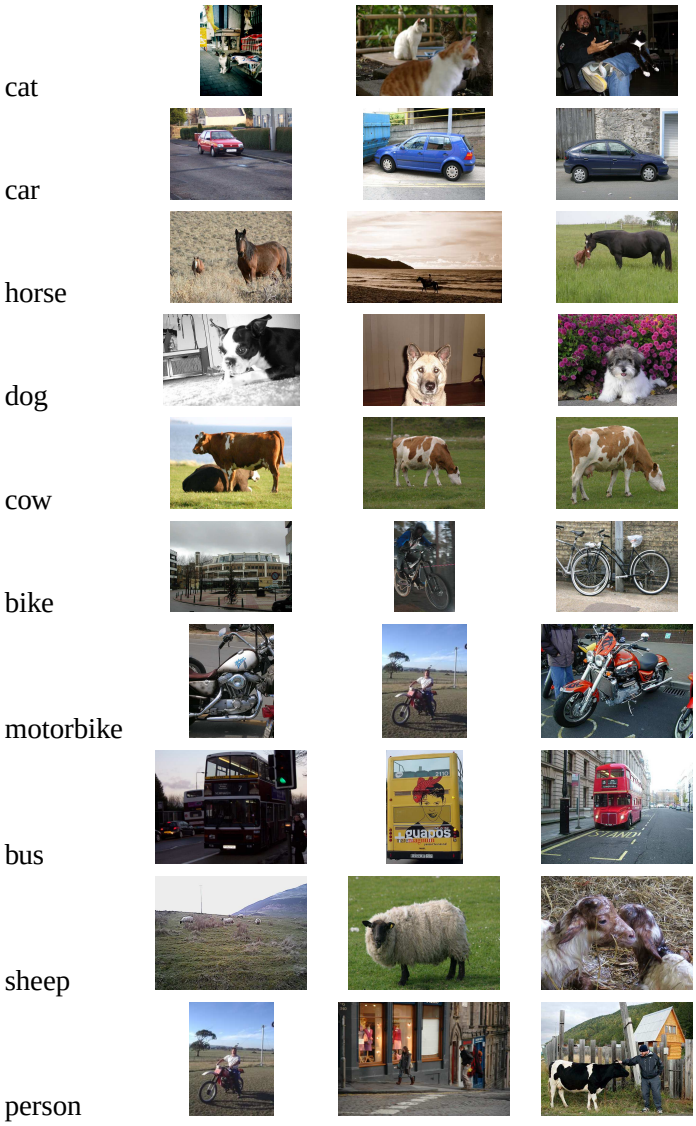
3.2.4 UIUC Cars

The UIUC cars database[4] was first introduced by Agarwal et al. in [3]. It contains 1050 training images (550 car and 500 non-car images) and 170 single-scale test images as well as 108 multi-scale test images. The training images are quite small (100x40) and quite roughly quantized, the test images are a bit bigger and may contain several cars. All images are in gray scale.

3.2.5 PASCAL Visual Object Class Challenge (2005/2006)

These benchmarks were proposed in 2005[5] and 2006[6] by the PASCAL (Pattern Analysis, Statistical modelling and ComputationAl Learning) network. In both challenges, two different kinds of tasks had to be solved: one was to predict the presence/absence of a class member in the test image, the second was to additionally draw a bounding box around the recognized objects (localization). There are no extra background images, but all other images from the database not containing the object form the counter class. In 2005, 4 object categories had to be distinguished (motorbikes,bicycles,people,cars), in 2006 there were 10 object classes (bicycle, bus, car, motorbike, cat, cow, dog, horse, sheep, person). For both challenges, many reference results are available. The images in the individual categories of VOC2005 were partially taken from other databases (Caltech, TuGraz, UIUC etc.), the VOC2006 data includes some images provided by Microsoft Research Cambridge and ”flickr”.

[4] http://l2r.cs.uiuc.edu/ cogcomp/Data/Car/
[5] http://www.pascal-network.org/challenges/VOC/voc2005/index.html
[6] http://www.pascal-network.org/challenges/VOC/voc2006/index.html



3.2.6 ImageCLEF2006 Object Annotation Task

This dataset is not as widely known as the previously described ones, and only two reference results exist. It is a quite hard database with 21 classes (ashtrays, backpacks, balls, banknotes, benches, books, bottles, calculators, cans, chairs, clocks, coins, computer_equipment, cups_mugs, hifi_equipment, knives_forks_spoons, mobilephones, plates, sofas, tables, wallets). The images presented should be automatically labelled with the right key word, i.e. classified correctly. The database and further information is available here[7].

3.2.7 MUSCLE Animal Images

Within the MUSCLE (Multimedia Understanding trough Semantics Computation and Learning) campaign, a dataset consisting of different animal categories was produced. The initial version proved too difficult, so

[7] http://www-i6.informatik.rwth-aachen.de/ deselaers/imageclef06/nonmedaat.html

a new database containing 262 images in 9 classes was released (cheetah, cougar, coyote, dear, goat, horse, leopard, lion, tiger). Results with this database are to await.

4 Object Class Recognition Using Patches

A recently popular method to deal with object class recognition is to use local information extracted at various points or areas in the image. Local patch based approaches have shown to have benefits over global methods: they are capable of modelling the variability in object appearance as well as the shape and can cope with occlusions. Both the kind of local information extracted (features) and the exact positions at which these are acquired can differ tremendously in different methods.

Methods based on locally acquired features are named differently in literature. The most common notion is "patch based" [63], however also the term "fragment based" [70], and sometimes "part based" is used. However, the term "part based" is mainly related to more semantic entities, e.g. arms, legs, body and head for a human being, where the first two are rather related to primitive extracts acquired from an image.

The use of local information has many advantages. We try to summarize them in the following list:

- **Reduction of the amount of data to be processed**
  Typically, the number of points where local information is extracted is significantly less than the number of pixels in the image. Information is either extracted at points "where something happens", so called "interest points" (see Section 4.2), a number of random points [39], points from a fixed size grid [12] or combinations of these. Also, features can be extracted from areas [53, 40].
- **Avoidance of segmentation**
  The objects to be recognized do not have to be segmented prior to recognition, in contrast, some patch based approaches even deliver a segmentation of the recognized objects [31].
- **Robustness to background clutter**
  This item is related to "avoidance of segmentation". When using local information, the classification step should ideally only consider parts that have a strong indication for the object itself, information from the background should be ignored ideally.
- **Robustness to occlusion**
  In many real world scenes, objects to be recognized are partially occluded. Global methods that require, e.g., the outline of an object, fail at this point. Patch based approaches have shown to cope well here, since the local information acquired at one point is not affected by other, occluded parts of the object.
- **Robustness to variability in object shape**
  When dealing with visual object classes, we have to cope with variability in the object configuration. Since the extraction of local information at one point is not affected from object parts at other locations, we gain robustness. We detach shape and appearance information, and can model them separately, as already proposed by Fischler and Elschlager [20].

On the other hand, the use of local patch information has also disadvantages:

- **Miss of relevant parts/structures of the image**
  When using an interest point/covariant region detector, there always is the danger that relevant parts of the object are missed. Later stages, that rely on these parts are likely to fail then. This is especially problematic if the interest point detector emits only few interest points, like the Harris/Hesse-Laplace or the DoG detector (see Section 4.2).

- **Loss of spatial coherence of the parts**
  If the location where the patches were extracted gets discarded as, e.g., in the "bag of features" approaches (see Section 4.7), we loose information. Some parts might only be discriminative within a geometric configuration.

These disadvantages have to be attenuated or canceled in order to achieve superior performance.

4.1 Basic Principles

Current patch based approaches for the recognition of visual object classes consist of several main steps, which are depicted in Figure 4. This is basically the common pattern recognition scheme, where feature extraction and learning is modelled in two steps.

- **Determine location and area of feature extraction**
  Since our premise is to use local information, we first have to determine where this local information should be extracted. Section 4.2 deals with this in detail.
- **Type of features to be extracted**
  A variety of features can be extracted from local areas. Ideally, they are robust to illumination changes as well as noise and capture the properties of the area they are extracted from well. In Section 4.3, we describe some feature extraction methods.
- **Learning**
  In order to describe the object class, we have to learn what is characteristic for it. We present training data to the system, either in a supervised, weakly supervised or unsupervised manner. Depending on whether a discriminative or generative method is applied, we obtain an object model or a decision function for the classification step. This step is where most approaches differ.
- **Classification/Localization**
  New images presented to the system for classification typically undergo the same interest point/area and feature extraction procedure as the training images. Then they are classified using the learned functions or object models form the database. Here again a variety of different methods are available, ranging from simple nearest neighbor techniques [12] to more advanced techniques like SVMs [72] or boosting procedures [53].
- **Validation/Tests**
  In order to judge the quality of the huge amount of procedures presented, they must be tested. A variety of benchmarks and reference databases exist, as could be seen in Section 3.2. Different standard measures like ROC (Reciever Operator Statistics) curves, PR (Precision-Recall) graphs or EER (Equal Error Rates) values can help to make the results more comparable.

4.2 Location of Feature Extraction

In literature, very diverse methods exist to determine where to extract features for object classification. Typically, features are extracted at so called interest points, however, the exact meaning of "interest point" differs from author to author. So Agarwal et al. [3] defines them to be *"points that have high information content in terms of the local change in signal."*, Cordelia Schmid et al. [56] as *"points where a signal changes two dimensionally"* or Loupias et al. [36] just as *"points where something happens in the signal at any resolution"*.

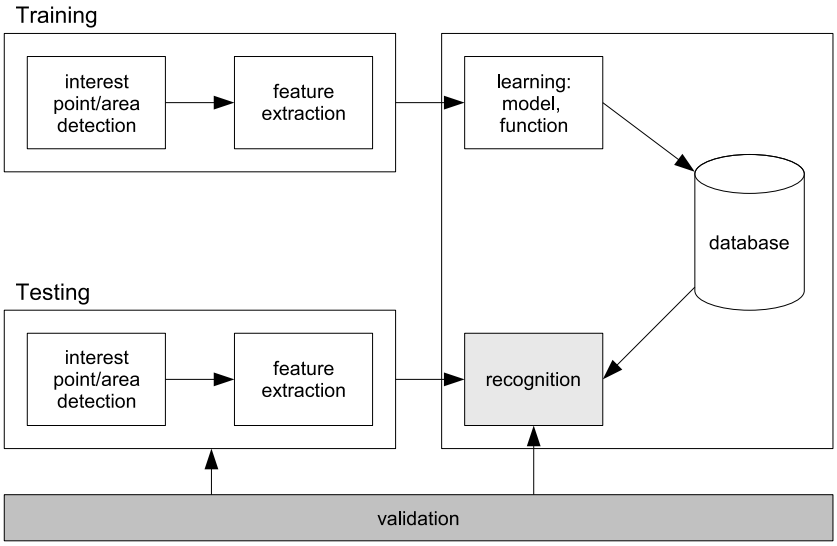


Figure 4: General scheme of object recognition using patches

Other authors like Reisert actually mean “specific part detectors” if they mention “interest point detectors”, since they do not detect any part where something happens, but only parts that have specific properties.

Interest point detection has a long tradition in classic computer vision for finding point correspondences to reconstruct 3D scenes from 2D views. There exist a lot of evaluation papers that try to judge the quality of interest point detectors, e.g. [58, 57, 47, 45]. The evaluation criteria are mainly repeatability (i.e. robustness against varying imaging conditions like viewpoint, scale, illumination changes) and information content. Repeatability however, is not necessarily a suitable measure when dealing with different objects, since direct point correspondences can not be established. Mikolaiczyk et al. [41] also evaluated interest point detectors (and features) in terms of object class recognition.

Scale invariance is also an issue if we deal with arbitrary images, so not only points, but rather areas like boxes or circles should get extracted from the images. Methods to achieve robustness to scale changes include finding function maxima in scale space [35] or the calculation of the entropy in an area [26].

Even affine distortions are considered by some detectors, they deliver ellipses [27, 43] or parallelograms [67]. These detectors are also referred to as “covariant region detectors”, since they change in a way consistent with the image transformation, where as an “invariant detector” would remain unchanged if we consider the strict meaning of the word.

Many authors, e.g. Mikolajczyk et al. [41] or Oplet et al. [51] suggest that different detectors should be used complementary, as they have different properties (some detect edge like, some corner like structures), so more information can be captured. Others like Deselaers et al. [12] state that the exact choice of the interest point detector is not important, as long as enough interest points are extracted. Even very simple means for locating interest points might be sufficient, e.g., just taking locations with high local grayvalue variance or entropy. Maree and Geuts [39] even go further: they just use a sufficiently large number of random points.

Homogeneous regions also carry information that can be beneficial for object class recognition, and thus some detectors search for regions with similar properties. The MSER detector (Maximally Stable Extremal

Regions) [40] or the IBR detector (Intensity Based Regions) [68] are a examples for that. These regions might have any form and are not restricted to a specific geometric layout.

In the following, we give a list of interest point/covariant region detectors, to adumbrate the varietey of them:

1. **(Classic) Interest point detectors**
    - Harris corner points [24]
    - Förstner point detector [22]
    - Wavelet based salient points [36]
    - Complex filters (Marco)
2. **Covariant region detectors**
    - Harris-Laplace/Harris-Affine [43]
    - Hesse-Laplace/Hesse-Affine [43]
    - MSER (Maximally Stable Extremal Regions) [40]
    - Laplacian of Gaussian (LoG)
      citeLindeberg1998
    - Difference of Gaussian (DoG) [38]
    - Intensity Based Regions (IBR) [68]
    - Edge Based Regions (EBR) [68]
    - Kadir & Bradey salient regions (based on entropy)[27]
3. **Other approaches for interest point/region detection**
    - Random points [39]
    - Grid (sparse, dense) [12]
    - Gradient magnitude [46]

It is generally noticed that some feature detectors deliver a very sparse representation, so that subsequent steps in the object recognition chain may suffer from that. So typically, many feature detectors are run, not only to get different types of interest points, but also to get more of them. Some authors also start to use exhaustive sampling of the image now, e.g. Fergus et al. [19].

4.3 Types of Features

The pixel values can not be compared directly, since they might undergo the variety of transformations mentioned above. In order to describe image structures, so called “features” get extracted, that should be robust to at least some of the imaging conditions.

Their complexity can vary from pure intensities (gray values) to sophisticated descriptors like SIFT (Scale Invariant Feature Transform) [37]. As already described in Section 3.1, an image might suffer from several imperfections like noise and non standard illumination. Features should either be robust against that or must be normalized, e.g. by color normalization, illumination normalization or scale normalization techniques. This is especially important for object class recognition from arbitrary images, since we have no knowledge about the recording conditions.

In this work, we mainly deal with local features. They are calculated from a relatively small, bounded region, that was acquired as described above. Global features on the contrary involve the whole object or even the entire image.

A huge variety of different types of features have been proposed for object recognition and image retrieval. Some of them were evaluated in [44]. As expected, different features are suitable for different tasks and objects. Opelt et al. [53] even propose to calculate many different features and then let the classifier decide which features to choose (boosting).

We just list some of the more commonly used:

1. **Texture features**
   tamura texture features, wavelets, gabor filters, steerable filters, (invariant) moments, jet features, local binary patterns (LBP), SIFT, SURF, GLOH, gray value invariants (monomial, relational kernels), cooccurrence features, MSA, PCA-SIFT
2. **Color/Intensity features**
   gray values (possibly dimensionally reduced: PCA, LDA, BDA), color histograms (important: color space, illumination, color constancy)
3. **Shape features**
   edge orientation histograms, line features, gradients (magnitude, orientation), shape context Fourier descriptors
4. **Others**
   gaussian derivative filters, differential invariants, complex filters, cross correlation of sampled pixel values, spin images

4.4 Learning

Our goal is to determine whether an object is present in an image or not. In order to decide that, pattern recognition offers two fundamentally different methods: generative or discriminative learning.

> **Generative methods** learn probabilistic object models $P(\text{input})$. For this, only the data of the current class is necessary. Typically, distributions for $P(\text{object present})$ and $P(\text{object absent})$ are learned, where $P(\text{object absent})$ is build from arbitrary background images or images containing other objects. For decision, typically a likelihood ratio test of the two choices is performed to classify new images. When adding a new object category, the old ones are not affected.
>
> Advantages of generative methods are that they can handle missing or partially labelled data and new classes can be added easily. Moreover, they can handle compositionality, where standard discriminative models need all possible object variations in order to learn the decision function [71].
>
> **Discriminative methods** a decision function is learned directly, i.e. $P(\text{class}|\text{input})$, e.g. by regression techniques. The decision function might be, e.g., be a parametric model, where the values of the parameters are inferred from a set of labelled training data, such as a neural network or an SVM.
>
> Advantages of discriminative methods are that they are usually very fast and are expected to be more accurate than generative methods, since they optimize the decision function directly [71].

Learning methods also differ in the amount of supervision they need for training. Strongly supervised methods typically need segmented objects or manual labels of specific object parts. If a huge variety of object classes is to be learned, this is not desirable. Weakly supervised methods only require the class labels per image, not where exactly the object is located. Unsupervised learning does not even require image labels.

Since for discriminative methods, the goal is a direct minimization of the classification error, the performance might be superior, however, when when training data is limited, generative approaches might be preferable, as shown by Ng and Jordan [49].

Often, generative models and discriminative learning are intermixed, e.g. [4, 23], in order to benefit from the two worlds.

4.5 Feature Sets

One important property of patch based approaches is that we do not have to deal with individual features or feature vectors describing the object, but with feature sets, which require some further processing or the modification of known pattern recognition techniques. For example, if we want to use an SVM on feature sets, we need specific set kernels for that [72].

The features are typically not ordered in the sets, and direct correspondences between features of two images can not be established easily. A lot of features in a particular set might be doubled, since they occur at many places of the object/image, or superfluous/erroneous, since they come from the background.

The lack of correspondence between the features of images can be handled two fold:

> **Ignore the lack of correspondence**: This leads to so called "bag of features" approaches, which are basically histograms of features. They show generally good performance, but localization can not be done, since all spatial information is lost during histograming.
>
> **Identify and correspond features**: In this approach, we are looking for specific "parts". Hillel et al. [4] define a part to be "*an entity with a fixed role (probabilistically modelled)*". Parts can either be something that corresponds to human perception like eyes, noses, wheels, more abstract pieces like an arbor or any other pattern. Whenever supervised learning is done, humans tend to identify semantic parts, however they do not necessarily mean that they are the most discriminative ones. E.g., the hair line is a good and stable part for the recognition of faces, however, people would not select this part at first sight. The number and type of parts vary greatly for different approaches, e.g. Fergus [18] constellation model has 3-7 parts where Agarwal et al. [2] use a codebook of 270 parts.

When we have detected specific object parts in an image, feature relations (mainly location and scale relations) provide a powerful cue for classification. The problem here is to find discriminative and geoetrically stable parts of the object reliably.

4.6 Clustering

A way to cope with the diversity and size of the feature sets is to cluster them, i.e. to group them according to some criteria and then to use the cluster centers only. In this way, we gain robustness to small variations in the patch. Typically, a similarity measure is applied for grouping the patches, e.g. (normalized) cross correlation or a Minkowsky norm. This leads to a more abstract, soft description of parts. Instead of the features themselves, the cluster means or some other representations are used for the different cluster members.

Clustering can also used as a mean for part selection [18]. When we cluster features from training images all containing the object, we assume that clusters with patches originating from the object should be big, where the background patches should end up in smaller clusters, since they should have arbitrary appearance. However, this is only true when the object is photographed in totally diverse scenes. Usually, the background

is also somehow similar (streets, office), so bigger clusters might as well be from the background. Maybe this information still helps for classification, e.g., if we want to discriminate cars on roads against office backgrounds (or the similar), then the street patches would also be discriminative (and it is to suspect that exactly this is the case in many object recognition systems.

Clustering can be seen as unsupervised learning, since initially, no labelling of the data is done. Depending on the clustering algorithm, we might obtain different clustering solutions, some of which might be more suitable than others for object class recognition. Even for the same algorithm, the exact solution can differ, because of random initialization and local minima.

There is a variety of different clustering algorithms, a good overview about them can be found in [64]. Cluster algorithms used for object recognition mainly fall into one of the two categories:

1. **Hierarchical clustering**

   Divisive clustering The data is first regarded as a whole, then it is split consecutively in smaller parts. Examples for that is e.g. the Linde-Buzo-Gray [34] algorithm. This approach is followed by [12, 13] as an example.

   Agglomerative clustering Here, initially all data entries are regarded as single clusters, and they are grouped with the most similar clusters in the following steps, until all data is grouped. This procedure can be visualized by dendrograms. To obtain individual clusters, the tree is "cut", so that either a certain number of clusters emerge, or the similarity of the clusters is above some threshold. A disadvantage of this approach is that very small clusters might occur, and this did not prove useful for object class recognition. Moreover, we have to deal with large time ($O(N^2 \log(N))$) and space ($O(N^2)$) complexity. As an advantage, we get visually very compact clusters. Methods falling into this category are applied, e.g., by [2, 29].

2. **Clustering based on function optimization**
   Here, a function J gets minimized, e.g. the distance of the data entries to the cluster centers. Typically, the number of clusters has to be given. How to select the ideal number of clusters is not always clear. It is commonly determined experimentally. Examples are k-means and EM clustering. For our experiments, clusters obtained by k-means were already well suited, they performed better than the agglomeratively clustered ones for cluster membership histograms. [74, 61] use clustering techniques of this kind.

The clusters obtained in this way are referred to as "visual codebook" or "visual words", depending on the background of the authors. Again, different strategies can be followed for the construction of the clusters: one is to obtain very specific codebook entries that describe one class particularly well, an other is to obtain very generic ones, so that the codebook might be used for many object classes. This leads to the question whether there should exist specialized codebooks for each object class or if there are generic ones where clusters (i.e. object parts) are shared by different object classes [66, 42]. Since clustering is usually very expensive, it would be desirable to have one general codebook that can be used for all classes, so that the addition of a new classes does not require new clustering in order to build a new dictionary or a complete rebuild of an existing codebook.

Recently, Miklolajczyk et al. [42] introduced a new mixed clustering approach, where first k-means clustering and then agglomerative clustering is performed in order to be able to deal with hundreds of thousands of feature vectors.

4.7 Object Geometry

The spatial layout of object parts is a powerful cue for recognition. Despite the fact that bag of word approaches often perform very well in two class object recognition tasks, the geometric relation of parts gives beneficial hints, especially in a multi class setting. Bag of word approaches perform well if the pure existence of certain structures give a strong indication whether an object is present or not. However, an object is more than the pure collection of parts, and some parts might become discriminative in a specific constellation only. So it is useful to study how object geometry might be modelled and used for recognition.

One difficulty is that for 3D objects in the real world, the geometric configuration of parts and sometimes even worse the parts visible might differ tremendously. One approach is to restrict the modelling of object geometry to a certain view and use several of them to recognize the object, an other one is to learn a true 3D model. Typically, the former approach is taken, where sometimes already transitions between views are learend [65].

A vital part for treating structure is that specific parts have to be identified. These parts can be strong ones with a semantic meaning like eyes, wheels and so on, or rather soft ones like edges, bars, corners, that might match at many positions, but where the distributions are known relative to an object reference point. Several possibilities exitst to decide which object parts to choose:

- Selection of patches by hand, e.g. by clicking on them [6]
- Selection of patches by exhaustive search and test on a validation set [73]
- Selection of patches directly by a classifier [53, 16].
- No selection at all, maybe reduction of parts by clustering if applicable [2]

Popular methods modelling the geometry of objects are variants of the “star model”, i.e. object parts in relation to a center point. Examples for this method are Leibe et al. [29], Fergus et al. [19] or Shotton et al. [59]. Another possibility is the “constellation model” [18] which models the joint probability of all parts to another. This can be seen as a new variant of the “parts and structure” approach already proposed by Fischler and Elschlager [20] in 1973, where objects are modelled as a collection of parts that are connected by springs.

5 Comparison of Systems

The recognition of visual object classes is a very active research field, and every year, a variety of new papers dealing with this topic are published. We want to give a small overview about approaches that had great impact on research in this area and describe current state of the art techniques.

The properties of most of these approaches are summarized in Table 1, classification error rates on standard datasets are listed in Table 2, 3, 4 and 5.

In the following list, which is only comprised of a fraction of the published literature on this topic[8], a huge variety of different methods were proposed to deal with the task of visual object class recognition. Typically, the systems make exhaustive use of machine learning techniques like EM, SVMs or Ada-boost, but also traditional Bayesian approaches are frequently used or intermixed. The approaches can be roughly divided into methods using the geometric distribution of parts or neglecting them, or in generative as well as discriminative methods. Current research trends tend to combine several previously proposed methods, e.g. different interest point detectors, different features or different matching strategies, which makes it difficult to judge

[8] the selcetion was based on what was considered to be important by the authors and may well be biased

the overall performance of the individual components, since the interplay and the fine tuning of the different parts becomes more and more important.

Burl, Weber, Perona [7]	1998
A Probabilistic Approach to Object Recognition Using Local Photometry and Global Geometry	
Ipts:	Manual selection of candidate parts, eyes, nosetip, mouth corners; correlation
Features:	Dominant local orientation
Distance:	Correlation
Classifier:	Likelihood ratio
Part detection via matched filtering, then a probabilistic shape model is applied (joint probability of parts) Translation, rotation and scale invariance is achieved by using relative positions to reference points.	

Weber, Welling, Perona [74]	2000
Unsupervised Learning of Models for Recognition	
Ipts:	Förstner detector
Features:	Gray values, gradients
Distance:	Normalized correlation
Classifier:	Likelihood ratio
Förstner interest points get extracted, then features calculated and clustered with k-means (100 clusters, only features from the positive class are used). Clusters with less than 10 members are removed, also clusters that are similar to others after a small shift. A number (3-7) of distinctive parts get selected, according to their classification performance on a validation set. For classification, the joint probability density of the detected part locations is evaluated.	

Agarwal, Roth et al. [3, 2]	2002, 2004
Learning a Sparse, Part Based Object Representation	
Ipts:	Förstner detector, sqare patches 13x13 pixel
Features:	Gray values
Distance:	Normalized cross correlation
Classifier:	SNOW (Sparse Network of Winnows)
Features are clustered in an image and the occurrence of cluster members at a specific spatial relationship is coded in a binary vector. As a classifier, winnow are used. For localization, a sliding Windows approach is used to calculate a classifier activation map, i.e. the probabilities, that an object at a certain location is present. Later, a multiscale approach was also proposed.	

Dorko, Schmid et al. [15, 16]	2003, 2005
Object Class Recognition Using Discriminative Local Features	
Ipts:	Kadir & Bradey, Harris-Laplace, Harris-affine, patches get normalized (geometry, scale, direction)
Features:	SIFT
Distance:	Gaussian Kernel Density
Classifier:	Number of "activated" part classifiers above threshold
Extraction of scale and affine covariant parts, calculation of SIFT features, clustering of the features with GMM, each Gaussian represents a cluster. Part classifiers are built (NN with Gaussian kernel density) and discriminative parts are found with two criteria: classification likelihood or mutual information. For classification, the n most discriminative object parts are used and the final decision is done whether the number of "activated" positive part classifiers is above a certain threshold, which is determined for each class.	

Fergus, Perona, Zisserman [18]	2003
Object Class Recognition by Unsupervised Scale-Invariant Learning	
Ipts:	Kadir & Bradey detector
Features:	Gray values, dimension PCA reduced (10D)
Distance:	Gaussian
Classifier:	Likelihood ratio
About 30 Kadir & Bradey regions get extracted and normalized to 11x11 pixels. The approach is similar to [74], however, here also the part appearance as well as the relative scale is modelled. All parameters of the model are learned via EM, even the selection of parts. A hypothesis vector assigns the detections to the previously learned parts or marks them as hidden. Classification is done using the likelihood ratio considering shape, appearance, scale and detector/occlusion statistics. Disadvantages are the long training and classification times, since a huge number of parameters has to be learned and part mappings have to be found. For this, all possible configurations of the detected parts are evaluated.	

Leibe, Schiele [31]	2003
Interleaved Object Categorization and Segmentation	
Ipts:	Harris detector
Features:	Gray values
Distance:	Normalized correlation
Classifier:	Hough like voting scheme
A codebook of object parts is generated using agglomerative clustering and normalized grayvalue correlation. For each codebook entry and object class, probabilities for object centers are calculated from a training set. Segmentation masks are stored for each codebook entry. For classification, a hough like voting scheme is applied, with that, probable object centers can be found. Using the backprojeckted hypothesis, a refined sampling can be done to get an improved hypothesis. The images can also be segmented using previously learned segmentation masks.	

Carbonetto, Dorko, Schmid [8]	2004
Bayesian Learning for Weakly Supervised Object Classification	
Ipts:	Harris-Laplace, Kadir & Bradey, LoG, DoG, Harris-affine, random selection
Features:	SIFT
Distance:	Gaussian kernel density
Classifier:	Probit link classifier
This works deals with object class recognition as data association problem: features form training images may contain the object or the background, task of the classifier is to reveal their affiliation. A probit link classifier is used for each patch, the parameters are learned by a MCMC (Markov Chain Monte Carlo) algorithm. The sum of the label probabilities for each patch in the image lead to a decision.	

Leibe, Schiele [33, 32]	2004
Interleaved Object Categorization and Segmentation, Scale Invariant Object Categorization Using a Scale-Adaptive Mean Shift Search	
Ipts:	Harris detector, DoG detector
Features:	Gray values
Distance:	Normalized Correlation
Classifier:	Hough like voting scheme
Improvements to [31]: The use of MDL (Minimum Description Length) for multi object recognition, Mean Shift search for fast maximum search in the Hough accumulator array and scale invariant interest point detection using the DoG detector.	

Csurka, Dance, Fan, Willamowski [10]	2004
Visual Categorization with Bag of Keypoints	
Ipts:	Harris affine
Features:	SIFT
Distance:	SVM: linear kernel
Classifier:	Naïve Bayes, linear SVM
SIFT features get extracted at geometrically normalized Harris affine patches. These are then clustered using k-means clustering and histograms of cluster memberships (“bag of keypoints”, “bag of words”). Classification is done by Naïve Bayes and SVMs, where the latter outperforms the former. All spatial relations are ignored in this approach.	

Torralba, Murphy, Freeman [66]	2004
Sharing Features: Efficient Boosting Procedures for Multiclass Object Detection	
Ipts:	Exhaustive search
Features:	Gray values templates (dimension 2000) and spatial masks
Classifier:	Joint boosting
As features, 2000 random patches get extracted from training images of the 21 object classes together with spatial masks. They are used as a kind of matched filter. For training, a "joint boosting" approach is proposed. This means that from the pool of weak learners, the one is chosen that not only separates a single class from the background best, but all the selected classes. This weak learner is then added to the strong learners of the selected subset. Results on both toy data as well as on real world images show that it is beneficial to use the shared classifiers, which - in this case - corresponds to shared features, since one weak classifier means one specific feature.	

Bar-Hillel, Weinshall [4]	2004
Efficient Learning of Relational Object Class Models	
Ipts:	Kadir & Bradey detector
Features:	Normalized grayvalues, DCT transformed, (15D)
Classifier:	Boosting
In this approach, the appearance, location and scale of parts are considered. A Bayesian network is used to learn the dependencies of part locations and part scales. Not all relations of parts are used to another, but only to an object reference point. The models have an intermediate number of parts (60) and the parameters are learned using boosting. For classification, the probability of the feature sets belonging to the class are calculated over marginalization.	

Opelt, Pinz, Fusenegger [50, 52, 53]	2004, 2005
Generic Object Recognition with Boosting	
Ipts:	Harris-Laplace, Harris-affine, DoG, regions acquired by "similarity measure segmentation"
Features:	Diverse (subsampeled gray values, basic moments, invariant moments, SIFT, intensity distributions and invariant moments for regions)
Distance:	Diverse (Euclidean, Mahalanobis, etc.)
Classifier:	Boosting
The rationale behind this approach is that the performance of individual detectors, descriptors and distance measures might be category specific. So they should be all offered to the classifier which should select the best combination. For this, a boosting framework is proposed. Not only interest point/region features are used, but also segmented areas.	

Ulusoy, Bishop [71]	2005
Generative Versus Discriminative Methods for Object Class Recognition	
Ipts:	DoG
Features:	SIFT
Classifier:	Discriminative: softmax model in a linear network, generative: Gaussian mixture model
The authors compare a discriminative and generative approach for object class recognition in a weakly supervised framework. Images are not classified directly, but the patches in the image. Whenever a patch is labelled to belong to an object class, the whole image gets this label and the object is regarded to be present. For the generative model, they use Gaussian mixtures and learn the parameters with an EM-style algorithm, for the generative model they use linear as well as non linear networks and a softmax model. The results are tested using a cow/sheep database.	

Sudderth, Torralba, Freeman, Willsky [62]	2005
Learning Hierarchical Models of Scenes, Objects, Parts	
Ipts:	Affine Covariant Regions (prob. Harris/Hesse-Affine, MSER)
Features:	SIFT
Classifier:	Maximum likelihood
Objects are modelled as a set of parts with an expected appearance and position, in an object centered coordinate frame. The parameters of this model are learned via a Gibbs sampler, which uses a graphical model to analytically average over many parameters. The approach only works for images with roughly aligned objects, as in the Caltech 101 object database. In a nice graphic, specific parts and the distribution of their location in the image is shown. The parts were obtained by getting about 30 parts per image and clustering them to 32 clusters. In the second part of the paper, a graphical model is used to also model the scene the object is in, but this is rather sketched as an idea.	

Fergus, Perona, Zisserman [19]	2005
A Sparse Object Category Model for Efficient Learning and Exhaustive Recognition	
Ipts:	Kadir & Bradey, multi-scale Harris, curves
Features:	Normalized gradients, dimensions reduced via PCA
Classifier:	Bayes (Likelihood ratio)
Basically, the constellation model of [18] gets improved from a speed point of view in that for the geometric layout, not a full joint probability density is used any more. Instead, one specific landmark gets determined and the geometric configuration depends only on this. The landmark is assumed always to be present (no occlusion of this part). Also, 3 different types of detectors were used, besides the Kadir & Bradey detector also the multiscale Harris and a curve detector (linked Canny edges, broken at bitangent points). Patches are represented using normalized gradient intensity, the dimensions are reduced via PCA. The selection of parts is again done using a validation set. For testing, exhaustive search is proposed: all PCA basis vectors get convolved with the image for the first k PCA components, then for each model part an activation map can be computed. However, this is again quite expensive.	

Deselaers, Keysers et al. [12, 13]	2005, 2006
Discriminative Training of Patch Cluster Histograms	
Ipts:	Loupias interest point detector + regular grid, squared patches at fixed and variable scale
Features:	Gray values, PCA transformed
Distance:	Euklidean distance, symmetric KLD
Classifier:	Bayes with discriminative training
Features get clustered and histograms build from the cluster memberships. Using the histogram, a variety of classification methods get tested, e.g. global patch search and voting, nearest neighbor, Naïve Bayes, generative single Gaussian and discriminative training. The last one gets identified best, as it weights the feature clusters according to their discriminativity. In the second paper, the approach gets improved, i.e. patches get extracted at various scales and brightness normalization is performed by removing the first PCA coefficient of the gray value features. Also SVMs were tested as a classifier.	

Mikolajczyk, Leibe, Schiele [42]	2006
Multiple Object Class Detection with a Generative Model	
Ipts:	Dense sampling at gradients, Laplacian scale selection
Features:	SIFT features, dimension reduced to 40 via PCA
Classifier:	Likelihood ratio
Features are calculated from all points in the image where the gradient magnitude is above a certain threshold. The scale at these points gets determined via Laplacian scale selection according to [46]. For every feature, a geometry term gets determined coding the distance and relative angle of the object center to the interest point, according to the dominant gradient orientation and the scale of this interest point. SIFT features get calculated at these areas and the dimension reduced to 40 via PCA. A top-down bottom up clustering method is applied: first, the data is partitioned using k-means, then for the individual clusters agglomerative clustering is performed. A hierarchical tree structure for appearance clusters is build, which is used for efficient similarity computation. Classification is done in Bayesian manner computing the likelihood ratio. This test is done at local maxima of the likelihood function of the object being present. Some additional tests are applied to determine whether objects of different classes share similar clusters or overlapping objects exist. In this way, the location, scale and orientation of multiple objects can be determined.	

Method	# ipts/image	clustering	part select.	scale inv.	rot. inv.	localization	multiObj.	geometry
Deselaers et al. [12]	1000+300	div. LBG, 512/4096	disc training	no	no	no	no	no
Deselaers et al. [13]	500+300	div. LBG, 4096	disc. train	yes, manual	no	no	no	no
Agarwal et al. [3, 2]	8	agglom., avg. link	no	yes, manual	no	yes	no	yes
Dorko et al. [15]	100-300	GMM	LikRat/MuI	yes, detector	yes	pre-step	no	no
Carbonetto et al. [8]	3x100	no	MCMC	yes, detector	yes	pre-step	no	no
Burl et al. [7]	5	no	manual	yes	yes	yes	no	yes
Weber et al. [74]	150	k-means(100)	EM+valid	no	no	yes	no	yes
Fergus et al. [18]	30	no	EM+valid	yes, detector	no	yes	no	yes
Fergus et al. [19]	60	no	EM+valid	yes, detector	no	yes	yes	yes
Leibe et al. [31, 33, 32]	8269/16	agglom, avg.link(2519)	nein	yes	no	yes	yes	yes
Opelt et al.[52, 53]	many	SIFT: k-means(100-300)	boosting	yes, detector	no	no	no	no
Csurka et al. [10]	avg. 360	k-means (1000)	no	yes, detecor	yes	no	no	no
Mikolajczyk et al. [42]	$2.5 \cdot 10^5$	k-means + agglom.	no	yes, detecor	yes	yes	yes	yes

Table 1: Summary of features for different systems

Caltech datasets	motorbikes_side	faces	airplanes	cars_rear	leaves	UIUC cars_side
[2]						21
[12]	1.5	5.8	2.6			
[13]	1.1	3.7	1.4			
[14]	1.3	3.9	0.8			
[16]	0.5	0.46	1.25		1.08	
[8]	0.0		0.2			
[29]						2.5
[18]	7.5	3.6	9.8	9.7		11.5
[19](various)	2.7	9.7	6.3	2.3		
[52, 53]	5.7	0.0	2.5	0.0		0.0

Table 2: Error rates for different approaches on the Caltech datasets

Graz-01 DB	bikes	people
[16]	8.0	12.0
[8]	8.0	16.0
[53]	16.5	23.5

Table 3: Error rates on the Graz-01 datasets

Graz-02 DB	bikes	people	cars
[53]	22.2	18.2	29.5

Table 4: Error rates on the Graz-02 datasets

UIUC	cars side
[2]	21.0
[18]	11.5
[29]	2.5
[52, 53]	0.0

Table 5: Error rates on the UIUC dataset

References

[1] *Merriam-Webster's Collegiate Dictionary*. Merriam-Webster, 11th edition, 2003.

[2] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(11):1475–1490, November 2004.

[3] Shivani Agarwal and Dan Roth. Learning a sparse representation for object detection. In *Proc. ECCV*, 2002.

[4] Aharon Bar-Hillel and Daphna Weinshall. Efficient learning of relational object class models. *IJCV*, submitted.

[5] Hans Burkhardt. Grundlagen der Bildverarbeitung (Mustererkennung) WS 05/06. Lecture Notes.

[6] Michael Burl. *Recognition of Visual Object Classes*. PhD thesis, Calif. Inst. of Technology, 1997.

[7] Michael C. Burl, Markus Weber, and Pietro Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *Proceedings of the ECCV '98*, pages 628–641, 1998.

[8] Peter Carbonetto, Gyuri Dorko, and Cordelia Schmid. Bayesian learning for weakly supervised object classification. Technical report, INRIA Rhone-Alpes, Grenoble, France, 2004.

[9] Jacopo M. Corridoni, Alberto Del Bimbo, and Enrico Vicario. Image retrieval by color semantics with incomplete knowledge. *J. Am. Soc. Inf. Sci.*, 49(3):267–282, 1998.

[10] G. Csurka, L. Dance, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on statistical Learning in computer vision*, pages 59–74, 2004.

[11] T. Deselaers, D. Keysers, and H. Ney. Fire - flexible image retrieval engine: Imageclef 2004 evaluation. In *CLEF - Cross Language Evaluation Forum*, 2004.

[12] T. Deselaers, D. Keysers, and H. Ney. Discriminative training for object recognition using image patches. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition,*, volume 2, pages 157–162, San Diego, CA, June 2005.

[13] T. Deselaers, D. Keysers, and H. Ney. Improving a discriminative approach to object recognition using image patches. In *Proc. of DAGM*, 2005.

[14] Thomas Deselaers, Andre Hegerath, Daniel Keysers, and Hermann Ney. Sparse patch-histograms for object classification in cluttered images. In *DAGM 2006, Pattern Recognition, 26th DAGM Symposium*, Lecture Notes in Computer Science, page accepted for publication, Berlin, Germany, September 2006.

[15] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 634–639 vol.1, 2003.

[16] Gyuri Dorko and Cordelia Schmid. Object class recognition using discriminative local features. Rapport de recherche RR-5497, INRIA - Rhone-Alpes, February 2005. not published yet.

[17] John P. Eakins and Margaret E Graham. Content-based image retrieval - a report to the jisc technology applications programme. Technical report, Institute for Image Data Research, University of Northumbria at Newcastle, 1999.

[18] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–27, Madison, Wisconsin, 2003.

[19] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, June 2005.

[20] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computer*, 22(1):67–92, 1973.

[21] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Qian Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: the qbic system. *Computer*, 28(9):23–32, 1995.

[22] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centers of circular features. In *Proc. on Intercommission Conference on Fast Processing of Photogrammetric Data*, Interlaken, Switzerland, 1987.

[23] Mario Fritz, Bastian Leibe, Barbara Caputo, and Bernt Schiele. Integrating representative and discriminant models for object category detection. In *Proc. ICCV*, 2005.

[24] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Proceedings of The Fourth Alvey Vision Conference*, 1988.

[25] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. ISBN: 0521540518.

[26] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, November 2001.

[27] T. Kadir, A. Zisserman, and M Brady. An affine invariant salient region detector. In *Proceedings of the 8th European Conference on Computer Vision*, pages 345–457, Prague, Czech Republic, 2004.

[28] Daniel Keysers. *Modelling of Image Variability for Recognition*. PhD thesis, RWTH Aachen, 2005.

[29] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Proceedings of the Workshop on Statistical Learning in Computer Vision*, Prague, Czech Republic, May 2004.

[30] B. Leibe and B. Schiele. Analyzing contour and appearance based methods for object categorization. In *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR'03)*, Madison, WI, June 2003.

[31] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *Proceedings of British Machine Vision Conference (BMVC'03)*, Norwich, UK, Sept. 2003.

[32] B. Leibe and B. Schiele. Scale invariant object categorization using a scale-adaptive mean-shift search. In *In Proceedings of the 26th German Pattern Recognition Symposium (DAGM'04)*, Tuebingen, Germany, August 2004.

[33] Bastian Leibe. *Interleaved Object Categorization and Segmentation*. PhD thesis, ETH Zurich, 2004. PhD Thesis No. 15752.

[34] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *Communications, IEEE Transactions on [legacy, pre - 1988]*, 28(1):84–95, 1980.

[35] Tony Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116, 1998.

[36] E. Loupias, N. Sebe, S. Bres, and J-M. Jolion. Wavelet-based salient points for image retrieval. In *International Conference on Image Processing*, 2000.

[37] David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision*, pages 1150–1157, Corfu, Greece, September 1999.

[38] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[39] R. Maree, P. Geurts, J. Piater, and L. Wehenkel. Random subwindows for robust image classification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 34–40 vol. 1, 2005.

[40] J. Matas, O. Chum, M.Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. BMVC*, pages 384–393, 2002.

[41] K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1792–1799, 2005.

[42] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. In *Proc. CVPR*, 2006.

[43] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. ECCV*, volume 1, pages 128–142, 2002.

[44] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630, 2005.

[45] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, Accepted, 2005.

[46] K. Mikolajczyk, A. Zisserman, and C. Schmid. Shape recognition with edge-based features. In *Proc BMVC*, 2003.

[47] Krystian Mikolajczyk and Cordelia Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–8, 2004.

[48] Wolfgang Müller. *Design and Implementation of a flexible Content Based Image Retrieval framework - The GNU Image Finding Tool*. PhD thesis, Universtiy of Geneva, 2001.

[49] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Proc. NIPS*, 2004.

[50] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Generic object recognition with boosting. Technical Report TR-EMT-2004-01, EMT, TU Graz, Austria, 2004.

[51] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proc. ECCV*, 2004.

[52] A. Opelt and A. Pinz. Object localization with boosting and weak supervision for generic object recognition. In *Proceedings of the 14th Scandinavian Conference on Image Analysis (SCIA)*, 2005.

[53] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. Generic object recognition with boosting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(3):416–431, 2006.

[54] Jean Ponce, Svetlana Lazebnik, F. Rothganger, and Cordelia Schmid. Towards true 3d object recognition. In *Reconnaissance des Formes et Intelligence Artificielle*, 2004.

[55] Simone Santini, Amarnath Gupta, Arnold Smeulders, Marcel Worring, and Ramesh Jain. Content based image retrieval at the end of the early years. *Pattern Analyis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380, Dec 2000.

[56] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(5):530–535, 1997.

[57] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[58] Nicu Sebe, Qi Tian, Etienne Loupias, Michael S. Lew, and Thomas S. Huang. Evaluation of salient point techniques. In *CIVR '02: Proceedings of the International Conference on Image and Video Retrieval*, pages 367–377, London, UK, 2002. Springer-Verlag.

[59] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proc. ECCV*, 2006.

[60] Sven Siggelkow, Marc Schael, and Hans Burkhardt. Simba - search images by appearance. In *Proc. DAGM*, 2001.

[61] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman. Discovering objects and their location in images. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 370–377 Vol. 1, 2005.

[62] E.B. Sudderth, A. Torralba, W.T. Freeman, and A.S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1331–1338 Vol. 2, 2005.

[63] Alexandra Teynor, Esa Rathu, Lokesh Setia, and Hans Burkhardt. Properties of patch based approaches for the recognition of visual object classes. In *Proc. DAGM*, 2006.

[64] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition*. Academic Press, third edition, 2006.

[65] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, , and L. Van Gool. Towards multi-view object class detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, June 2006.

[66] Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *Proc. CVPR*, 2004.

[67] Tinne Tuytelaars and Luc J. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *Proc. BMVC*, 2000.

[68] Tinne Tuytelaars and Luc van Gool. Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision*, 59(1):61–85, 2004.

[69] S. Ullman and E Sali. Object classification using a fragment-based representation. In Seong-Whan Lee, Heinrich H. Bulthoff, and Tomaso Poggio, editors, *Biologically Motivated Computer Vision First IEEE International Workshop, BMVC 2000,*, number 1811 in Lecture Notes in Computer Science, pages 73–87. Springer, May 15-17 2000.

[70] S. Ullman, E. Sali, and M. Vidal-Naquet. A fragment based approach to object represenation and classification. In *Proc. of the 4th International Workshop on Visual Form*, Capri, Italy, 2001.

[71] I. Ulusoy and C.M. Bishop. Generative versus discriminative methods for object recognition. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 258–265 vol. 2, 2005.

[72] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *Proceedings of the International Conference on Computer Vision*, 2003.

[73] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *In Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition*, 2000.

[74] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *In Proceedings of the 6th European Conference of Computer Vision*, Dublin, Ireland, 2000.