Project/Team LEAR : Learning and Recognition in Vision

Cordelia Schmid

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team LEAR

Learning and Recognition in Vision

Rhône-Alpes

THEME COG

Table of contents

1. Team

LEAR has been part of the GRAVIR-IMAG laboratory, a Joint Research Unit of INRIA, the Centre National de Recherche Scientifique (CNRS), the Institut National Polytechnique de Grenoble (INPG) and the Université Joseph Fourier (UJF) till December 2006. Starting January 2007 it is part of the LJK laboratory.

Head of the team

Cordelia Schmid [Research Director INRIA (DR2), habilité(e)]

Deputy-head and scientific co-director

Bill Triggs [Researcher CNRS (CR1), habilité(e)]

Permanent researchers

Frédéric Jurie [Researcher CNRS (CR1), habilité(e)]
Hervé Jégou [Researcher INRIA (CR2), from 09/2006]

Faculty members

Roger Mohr [Professor and head at ENSIMAG, habilité(e)]
Laurent Zwald [Associate professor at UJF, from 10/2006]

Administrative assistant

Anne Pasteur [Secretary INRIA]

Postdoctoral fellows

Hakan Cevikalp [Techno-Vision project ROBIN, 08/2006-07/2007]
Vittorio Ferrari [EADS fellowship, 10/2005-11/2006]
Xiaoyang Tan [EU project CLASS, 09/2006-03/2007]
Joost Van de Weijer [Marie Curie fellowship, 04/2005-10/2007]
Jakob Verbeek [INRIA, 12/2005-11/2007]

Technical staff

Julien Bohne [MBDA grant, 10/2005-10/2006]
Matthijs Douze [EU project aceMedia, 01/2005-12/2007]
Benjamin Ninassi [Techno-Vision project ROBIN, 02/2005-02/2007]
Yves Gufflet [EU project CLASS, 05/2006-05/2007]

PhD students

Alexander Kläser [INPG, EU project CLASS, from 11/2006]
Ankur Agarwal [INPG, MENESR scholarship, 10/2004-05/2006]
Juliette Blanchet [UJF, MENESR scholarship co-supervised w. INRIA team MISTIS, from 10/2004]
Christopher Bourez [INPG, EU project CLASS, from 12/2006]
Charles Bouveyron [UJF, MENESR scholarship co-supervised w. INRIA team MISTIS, 10/2003-09/2006]
Christophe Damerval [UJF, MENESR scholarship co-supervised w. MOSAIC team of LMC, from 10/2004]
Navneet Dalal [INPG, EU project aceMedia, 10/2003-06/2006]
Matthieu Guillaumin [INPG, ENS Ulm scholarship, from 09/2006]
Diane Larlus [INPG, MENESR scholarship, from 10/2005]
Marcin Marszalek [INPG, Marie Curie project VISITOR, from 09/2005]
Eric Nowak [INPG, CIFRE scholarship from Bertin, from 02/2004]

MSc students

Alexander Kläser [Master Bonn-Klein-Sieg University, 02/2006-09/2006]
Hedi Harzallah [Master ENSI Tunis, 02/2006-01/2007]
Matthieu Guillaumin [Master ENS Ulm, 03/2006-08/2006]
Gagan Gupta [Master Indian Institute of Technology, 06/2006-06/2007]
Frank Moosmann [Master Karlsruhe University, 11/2005-04/2006]

Student interns

Loïc Février [ENS Ulm, 06/2006-07/2006]

Visiting scientist

Tinne Tuytelaars [KU Leuven, regular visits, 02/2006-03/2007]

2. Overall Objectives

2.1. Overall Objectives

LEAR's main focus is learning based approaches to visual object recognition and scene interpretation, particularly for image retrieval, video indexing and the analysis of humans and their movements. Understanding the content of everyday images and videos is one of the fundamental challenges of computer vision and we believe that significant advances will be made over the next few years by combining state of the art image analysis tools with emerging machine learning and statistical modeling techniques.

LEAR's main research areas are:

- **Image features and descriptors and robust correspondence.** Many efficient lighting and viewpoint invariant image descriptors are now available, such as affine-invariant interest points and histogram of oriented gradient appearance descriptors. Our current research aims at extending these techniques to give better characterizations of visual object and texture classes as well as 2D and 3D shape information, and at defining more powerful measures for visual salience, similarity, correspondence and spatial relations.
- **Statistical modeling and machine learning for visual recognition.** Our work on statistical modeling and machine learning is aimed mainly at making them more applicable to visual recognition and image analysis. This includes both the selection, evaluation and adaptation of existing methods, and the development of new ones designed to take vision specific constraints into account. Particular challenges include: (i) the need to deal with the *huge volumes of data* that image and video collections contain; (ii) the need to handle *rich hierarchies of natural classes* rather than just make simple yes/no classifications; and (iii) the need to capture enough domain information to allow *generalization from just a few images* rather than having to build large, carefully marked-up training databases.
- **Visual recognition and content analysis.** Visual recognition requires the construction of exploitable visual models of particular objects and of object and scene categories. Achieving good invariance to viewpoint, lighting, occlusion and background is challenging even for exactly known rigid objects, and these difficulties are compounded when reliable generalization across object categories is needed. Our research combines advanced image descriptors with learning to provide good invariance and generalization. Currently the selection and coupling of image descriptors and learning techniques is largely done by hand, and one significant challenge is the automation of this process, for example using automatic feature selection and statistically-based validation diagnostics.
- **Human detection and activity analysis.** Humans and their activities are one of the most frequent and interesting subjects of images and videos, but also one of the hardest to analyze owing to the complexity of the human form, clothing and movements. Our research in this area uses machine learning techniques and robust visual shape descriptors to characterize humans and their movements with little or no manual modeling. Particular focus lies on robust human detection in images and videos.

3. Scientific Foundations

3.1. Image features and descriptors and robust correspondence

Reliable image features are a crucial component of any visual recognition system. Despite much progress, research is still needed in this area. Elementary features and descriptors suffice for a few applications, but their lack of robustness and invariance puts a heavy burden on the learning method and the training data, ultimately limiting the performance that can be achieved. More sophisticated descriptors allow better inter-class separation and hence simpler learning methods, potentially enabling generalization from just a few examples and avoiding the need for large, carefully engineered training databases.

The feature and descriptor families that we advocate typically share several basic properties:

- **Locality and redundancy:** For resistance to variable intra-class geometry, occlusions, changes of viewpoint and background, and individual feature extraction failures, descriptors should have relatively small spatial support and there should be many of them in each image. Schemes based on collections of image patches or fragments are more robust and better adapted to object-level queries than global whole-image descriptors. A typical scheme thus selects an appropriate set of image fragments, calculates robust appearance descriptors over each of these, and uses the resulting collection of descriptors as a characterization of the image or object (a "bag-of-features" approach – see below).
- **Photometric and geometric invariance:** Features and descriptors must be sufficiently invariant to changes of illumination and image quantization and to variations of local image geometry induced by changes of viewpoint, viewing distance, image sampling and by local intra-class variability. In practice, for local features geometric invariance is usually approximated by invariance to Euclidean, similarity or affine transforms of the local image.
- **Repeatability and salience:** Fragments are not very useful unless they can be extracted reliably and found again in other images. Rather than using dense sets of fragments, we often focus on local descriptors based at particularly salient points – "keypoints" or "points of interest". This gives a sparser and thus potentially more efficient representation, and one that can be constructed automatically in a preprocessing step. To be useful, such points must be accurately relocalizable in other images, with respect to both position and scale.
- **Informativeness:** Notwithstanding the above forms of robustness, descriptors must also be informative in the sense that they are rich sources of information about image content that can easily be exploited in scene characterization and object recognition tasks. Images contain a lot of variety so high dimensional descriptions are required. The useful information should also be manifest, not hidden in fine details or obscure high-order correlations. In particular, image formation is essentially a spatial process, so relative position information needs to be made explicit, e.g. using local feature or context style descriptors rather than global moments or Fourier descriptors.

Partly owing to our own investigations, features and descriptors with some or all of these properties have become popular choices for visual correspondence and recognition, particularly when large changes of viewpoint may occur. One notable success to which we contributed is the rise of "bag-of-features" methods for visual object recognition. These characterize images by their (suitably quantized or parameterized) global distributions of local descriptors in descriptor space. (The name is by analogy with "bag-of-words" representations in document analysis. The local features are thus sometimes called "visual words"). The representation evolved from texton based methods in texture analysis. Despite the fact that it does not (explicitly) encode much spatial structure, it turns out to be surprisingly powerful for recognizing more structural object categories.

Our current research on local features is focused on creating detectors and descriptors that are better adapted to describe object classes, on incorporating spatial neighborhood and region constraints to improve

informativeness relative to the bag-of-features approach, and on extending the scheme to cover different kinds of locality.

3.2. Statistical modeling and machine learning for image analysis

We are interested in learning and statistics mainly as technologies for attacking difficult vision problems, so we take an eclectic approach, using a broad spectrum of techniques ranging from classical statistical generative and discriminative models to modern kernel, margin and boosting based machines. Hereafter we enumerate a set of approaches that address some problems encountered in this context.

- Parameter-rich models and limited training data are the norm in vision, so overfitting needs to be estimated by cross-validation, information criteria or capacity bounds and controlled by regularization, model and feature selection.
- Visual descriptors tend to be high dimensional and redundant, so we often preprocess data to reduce it to more manageable terms using dimensionality reduction techniques including PCA and its nonlinear variants, latent structure methods such as Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA), and manifold methods such as Isomap/LLA.
- To capture the shapes of complex probability distributions over high dimensional descriptor spaces, we either fit mixture models and similar structured semi-parametric probability models, or reduce them to histograms using vector quantization techniques such as K-means or latent semantic structure models.
- Missing data is common owing to unknown class labels, feature detection failures, occlusions and intra-class variability, so we need to use data completion techniques based on variational methods, belief propagation or MCMC sampling.
- Weakly labelled data is also common – for example one may be told that a training image contains an object of some class, but not where the object is in the image – and variants of unsupervised, semi-supervised and co-learning are useful for handling this. In general, it is expensive and tedious to label large numbers of training images so less supervised data mining style methods are an area that needs to be developed.
- On the discriminative side, machine learning techniques such as Support Vector Machines, Relevance Vector Machines, and Boosting, are used to produce flexible classifiers and regression methods based on visual descriptors.
- Visual categories have a rich nested structure, so techniques that handle large numbers of classes and nested classes are especially interesting to us.
- Images and videos contain huge amounts of data, so we need to use algorithms suited to large-scale learning problems.

As part of our work in this area, we maintain active links with both the statistics community, particularly via collaborations with the INRIA projects MISTIS and SELECT (formerly IS2), and the machine learning one, most notably via the EU project CLASS and the Network of Excellence PASCAL.

3.3. Visual recognition and content analysis

Current progress in visual recognition shows that combining advanced image descriptors with modern learning and statistical modeling techniques is producing significant advances. We believe that, taken together and tightly integrated, these techniques have the potential to make visual recognition a mainstream technology that is regularly used in applications ranging from visual navigation through image and video databases to human-computer interfaces and smart rooms.

The recognition strategies that we advocate make full use of the robustness of our invariant image features and the richness of the corresponding descriptors to provide a vocabulary of base features that already goes a long way towards characterizing the category being recognized. Trying to learn everything from scratch using simpler, non-invariant features would require far too much data: good learning cannot easily make up for bad features. The final classifier is thus responsible "only" for extending the base results to larger amounts of intra-class and viewpoint variation and for capturing higher-order correlations that are needed to fine tune the performance.

That said, learning is not restricted to the classifier and feature sets can not be designed in isolation. We advocate an end-to-end engineering approach in which each stage of the processing chain combines learning with well-informed design and exploitation of statistical and structural domain models. Each stage is thoroughly tested to quantify and optimize its performance, thus generating or selecting robust and informative features, descriptors and comparison metrics, squeezing out redundancy and bringing out informativeness.

One special case is detecting and recognizing humans, tracking their motions, and recognizing their activities. The importance of humans as subject matter and the complexities of their forms, appearances and motions warrant a special effort in this area. Our current research focuses on reliable detection and body-part labeling of humans in images and videos despite changes of viewpoint (from long shot to close up), lighting, clothing and pose. Future research will extend this to the classification and analysis of human actions.

4. Application Domains

4.1. Application Domains

A solution to the general problem of visual recognition and scene understanding will enable a wide variety of applications in areas including human-computer interaction, image retrieval and data mining, medical and scientific image analysis, manufacturing, transportation, personal and industrial robotics, and surveillance and security. With the ever expanding array of image sources, visual recognition technology is likely to become an integral part of many information systems. A complete solution to the recognition problem is unlikely in the near future, but even partial solutions in these areas enable many applications. LEAR's research focuses on developing basic methods and general purpose solutions rather than on a specific application area. Nevertheless, we have applied our methods in several different contexts.

Semantic-level image and video access. This is an area with considerable potential for future expansion owing to the huge amount of visual data that is archived. Besides the many commercial image and video archives, it has been estimated that as much as 96% of the new data generated by humanity is in the form of personal videos and images[1] and there are also applications centering on on-line treatment of images from camera equipped mobile devices (e.g. navigation aids, recognizing and answering queries about a product seen in a store). Technologies such as MPEG-7 provide a framework for this, but they will not become generally useful until the required mark-up can be supplied automatically. The base technology that needs to be developed is efficient, reliable recognition and hyperlinking of semantic-level domain categories (people, particular individuals, scene type, generic classes such as vehicles or types of animals, actions such as football goals, etc). In the EU FP6 project AceMedia we are currently developing methods that reliably find humans in still images and videos as well as methods for semi-automatic structuring of personal photo collections. In

[1] http://www.sims.berkeley.edu/research/projects/how-much-info/summary.html

the EU FP6 project CLASS we investigate methods for visual learning with little or no manual labeling and semantic-level image and video querying.

Visual (example based) navigation. The essential requirement here is robust correspondence between observed images and reference (map) ones, despite large differences in viewpoint. The reference database is typically also large, requiring efficient indexing of visual appearance. Both of these are core technology areas for our team. There are applications to pedestrian and driver aids and to autonomous vehicles including civilian (e.g. hospital robot) and aerospace and military ones. Our recent past project related to this area is the EU FP5 project LAVA (learning based methods and visual recognition applications suitable for use with mobile devices such as telephones with cameras).

Automated surveillance. This requires the reliable detection and recognition of domain classes, often in less common imaging modalities such as infrared and under significant processing constraints. Our expertise in generic recognition and in human detection and tracking is especially relevant here. Our current project with Bertin is on vehicle classification and on the evaluation of general object recognition techniques in such a context.

5. Software

5.1. Groups of adjacent contour segments

Participants: Vittorio Ferrari, Frédéric Jurie, Cordelia Schmid.

The PAS detection software is a freely available Linux executable for extracting and describing local contour features from images, http://lear.inrialpes.fr/software. It it quite portable and includes a README explaining fully how to use the software and the meaning of the various output files. The executable does not need matlab to run, thus circumventing potential license problems, and supports parallel processing of the same image directory by multiple computers. For information about kAS features, please refer to the related INRIA technical report [52] as well as section 6.1.1.

5.2. Histogram of oriented gradient object detection

Participants: Navneet Dalal, Bill Triggs, Cordelia Schmid.

As part of the European Union FP6 Integrated Project aceMedia we have developed a toolkit for detecting specific visual object classes such as humans, cars and motorbikes in static images. Although developed originally for human detection [4], the software implements a generic framework that can be trained to detect any visual class with a moderately stable appearance. The method has proven quite popular owing to its accuracy and its relative simplicity, with at least six academic or corporate research groups independently reimplementing it and more than 60 first-time downloads (http://lear.inrialpes.fr/software) since September 2005. The software is under copyright protection, registered at the Agence pour la Protection des Programmes (APP).

5.3. Extracting and describing interest points

Participants: Julien Bohne, Matthijs Douze, Frédéric Jurie, Cordelia Schmid, Bill Triggs.

Local descriptors [54] computed at affine invariant local regions [55] provide a stable image characterization in the presence of significant viewpoint changes. This provides robust image correspondence despite large changes in viewing conditions, which in turn allows rapid appearance based indexing in large image databases. Over the past several years we have been developing efficient software for this, http://lear.inrialpes.fr/software. Furthermore, in collaboration with Oxford, Leuven and Prague we designed a test setup which includes comparison criteria and a set of images containing representative scenes viewed under different transformations. This setup is available on the Internet (same address as above) and is currently used in the literature to evaluate new detectors and descriptors.

5.4. Image search demonstrator

Participants: Matthijs Douze, Cordelia Schmid, Bill Triggs.

Our image search technology takes as input a query image and outputs the most similar images in the database. The method extracts invariant local descriptors in each image, see previous paragraph, stores them in an efficient search structure, and measures similarity with a voting style approach.

This technology is used in two demonstrators. One is designed for interactive use. It finds the images containing any given object or scene element in a database containing 500 images in about a third of a second. The search object is defined by giving a sample image, for example from a webcam. Images can also be added to the database on the fly, allowing the system to be used for, e.g. vision based navigation. This method was demonstrated at the 2006 “Fête de la Science” and at the Forum 4i, Grenoble, 2006.

The second demonstrator focuses on efficient search methods for retrieval in larger databases. The current prototype (http://pascal.inrialpes.fr/dbdemo) is capable of finding an example image in a database of 50,000 images in a few seconds.

5.5. Signal processing and coding library

Participants: François Cayre [LIS], Vivien Chappelier [external contributor], Hervé Jégou [maintainer].

Libit is a C library for information theory and signal processing, http://libit.sourceforge.net. It extends the C language with vector, matrix, complex and function types, and provides some common source coding, channel coding and signal processing tools. The goal of libit is to provide easy to use efficient tools, and is mainly targeted at researchers and developers in the fields of coding or signal processing. The syntax is purposely close to that of other tools commonly used in these fields, such as MATLAB, octave, or IT++. Therefore, experiments and applications can be developed, ported and modified simply. Additional goals of the library include portability to many platforms and architecture, and ease of installation. Rather than trying to provide the latest state-of-the-art techniques or a large panel of specific methods, this library aims at providing the most general and commonly used tools in signal processing and coding. Among these tools are some common source models, quantization techniques, wavelet analysis, entropy coding, etc. As examples and to ensure the correctness of the algorithms with respect to published results, some test programs are also provided.

5.6. Datasets

Participants: Frédéric Jurie, Benjamin Ninassi, Navneet Dalal, Vittorio Ferrari, Cordelia Schmid, Bill Triggs.

Relevant datasets are important to assess existing recognition method. Furthermore, they allow to point out the weakness of existing methods and push forward the state-of-the-art. Datasets should capture a large variety of situations and conditions, i.e., include occlusions, view pose changes, bad illumination, etc. Benchmarking procedures allow to compare the relative strengths of different approaches, and providing a clear and broadly understood performance measure is therefore essential.

Today, there does not exist a standardized definition of what constitutes a good object detection/recognition system. There exists a clear need for sharing common datasets and metrics, and for introducing rigor in the datasets and benchmark procedures. One of the greatest challenges raised by benchmarking is the availability of shared test databases. These databases do not only need to contain thousands of images with associated ground truths, but they must be – as much as possible – royalty-free so they can be distributed.

We have recently been involved in creating several datasets, most importantly the *Robin dataset* funded by a Techno-Vision grant, see section 8.1.2 for details, but also our own research datasets the *INRIA humans and horses datasets*.

5.6.1. The Robin dataset

The Robin dataset consists of annotated images as well as performance metrics for multi-class object detection, object recognition and image categorization. The database includes six different datasets containing several hundred of annotated images which can be downloaded from the

Figure 1. Online image visualization tool.

ROBIN web site (http://robin.inrialpes.fr). A tool for visualizing and annotating images online with a web navigator is provided, see figure 1. Furthermore, the Robin dataset includes a set of carefully chosen metrics which satisfy the needs expressed by companies, competitors and evaluators, see http://robin.inrialpes.fr/robin_evaluation/downloads/ROBIN_metrics_v6.pdf for additional information. Competitions and benchmarking have started in November and will end in July 2007.

5.6.2. INRIA horses and human datasets

The INRIA human dataset was collected as part of research work on detection of upright people in images and video. The dataset is divided in two formats: (a) original images with corresponding annotation files, and (b) positive images in normalized 64x128 pixel format with original negative images.

The INRIA horse dataset was collected as part of our research on shape-based descriptors. It consists of 170 images with one or more side-views of horses, all of them annotated with bounding boxes, and 170 images without horses. Horses appear at several scales, and against cluttered backgrounds.

Both datasets can be downloaded from the Lear website at http://lear.inrialpes.fr/data.

6. New Results

6.1. Image descriptors and correspondence

Participants: Hervé Jégou, Frédéric Jurie, Cordelia Schmid, Bill Triggs, Christophe Damerval, Vittorio Ferrari, Eric Nowak, Joost Van de Weijer, Gyuri Dorkó [University of Darmstadt], Caroline Pantofaru [CMU], Matti Pietikäinen [University of Oulu], Tinne Tuytelaars [KU Leuven].

Our effort this year has concentrated on designing image descriptors for object classes. While descriptors designed for specific classes give excellent performance in the presence of significant changes of the viewing conditions, the question which descriptors are most appropriate to describe object classes is still open.

6.1.1. Groups of adjacent contour segments for object detection

We introduced a family of scale-invariant local shape features formed by groups of connected, roughly straight contour segments, and their use for object class detection [52], see figure 2. The pairs of adjacent contour segments (PAS) are able to cleanly encode pure fragments of an object boundary, without including

nearby clutter. Moreover, they offer an attractive compromise between information content and repeatability, and encompass a wide variety of local shape structures. We also define a translation and scale invariant descriptor encoding the geometric configuration of the segments within a PAS, making PAS easy to reuse in other frameworks, for example as a replacement or addition to interest points.

We demonstrate the high performance of PAS within a simple but powerful sliding-window object detection scheme, see figure 3. Through extensive evaluations, involving eight diverse object classes and more than 1400 images, we show that PAS substantially outperform interest points for detecting shape-based classes.

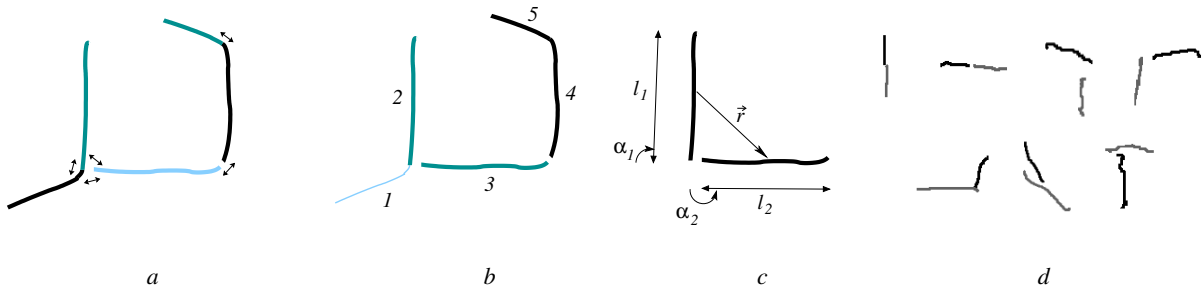


Figure 2. (a) Contour segment network edgel-chains connected at endpoints and junctions. (b) PAS = groups of two connected contour segments. (c) 5D PAS descriptor: orientations, relative lengths and location. (d) PAS codebook.

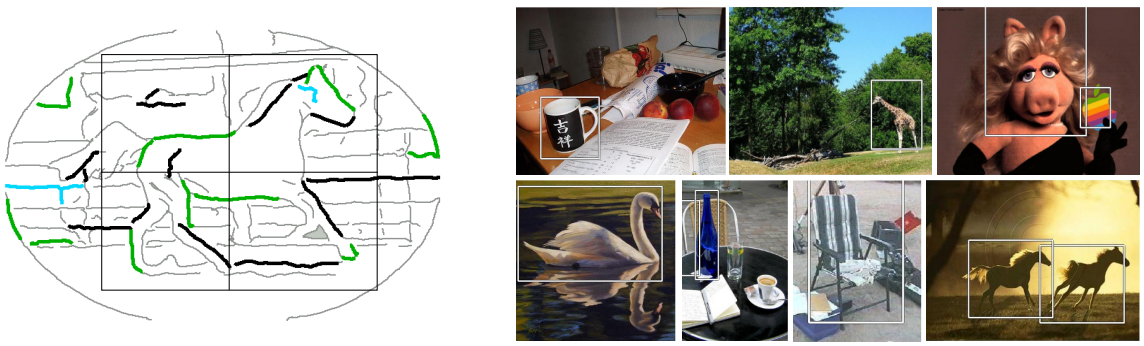


Figure 3. (a) Some of the PAS features as well as the tiled search window. (b) Example detection results.

6.1.2. Subsampling of dense patches

Bag-of-features representations have recently become popular for content based image classification owing to their simplicity and good performance. They basically describe the distribution of quantized local image regions. We recently proposed [41] to experimentally study the different aspects of this representation, including codebook size and creation method, histogram normalization method, minimum scale for feature extraction and feature sampling strategies. The later point is the main subject of our work. We showed that for a representative selection of commonly used test databases and for moderate to large numbers of samples, random sampling gives equal or better classifiers than the sophisticated multi-scale interest operators, see Figure 4 for examples. Although interest operators work well for small numbers of samples, the single most

important factor governing performance is the number of patches sampled from the test image and ultimately interest operators cannot provide enough patches to compete.

Figure 4. Examples of different sampling methods. (1) Harris-Laplace (HL) with a large detection threshold. (2) HL with threshold zero – note that the sampling is still quite sparse. (3) Laplacian-of-Gaussian. (4) Random sampling

6.1.3. Class-specific features for category-level recognition

The first steps in most recent class-level object recognition system consist of feature extraction, feature description, and clustering into a visual vocabulary. Our approach reduces this typical processing scheme to the bare essential. Rather than relying on 'designed' local feature detectors looking for e.g. corners or blobs, our system learns which features to use for a given classification problem, starting from densely sampled patches described by a robust, SIFT-like descriptor. This feature space is discretized, keeping the high dimensionality under control by storing only non-empty bins, using a table lookup. Feature selection is then performed simultaneously with the construction of a visual vocabulary, exploiting the learnt probability distribution and using a novel distance measure that takes the spatial structure of the SIFT-like descriptor into account, resulting in a class-specific, discriminative visual vocabulary. Experimental results on object classification and localization demonstrate the viability of the approach. This is joint work with T. Tuytelaars from University of Leuven.

6.1.4. Interest region descriptors: robustness to illumination changes

We propose a novel interest region descriptor [32] which combines the strengths of the well-known SIFT descriptor and the LBP texture operator, called the center-symmetric local binary pattern (CS-LBP) descriptor. Instead of using gradient orientation and magnitude based features as in SIFT, we employ CS-LBP features where each pixel is described by the relative gray levels of its neighboring pixels. If the gray level of the neighboring pixel is higher or equal, the value is set to one, otherwise to zero. These features are very robust to illumination changes.

The CS-LBP descriptor was compared to the SIFT descriptor within a recent test framework [55]. Our descriptor performed significantly better than SIFT in the presence of illumination changes. For the other test cases it performed better or about equal. Furthermore, our features are more robust on flat image areas, since insignificant gray level differences do not influence the thresholded results. It should also be noted that the CS-LBP descriptor is computationally simpler than the SIFT descriptor. This is joint work with M. Heikkilä and M. Pietikäinen from University of Oulu.

6.1.5. Image correspondence based on a contextual dissimilarity measure

Building on the *Video-Google* image retrieval setup [56], we have designed an enhanced scheme that provides better results than the state-of-the-art methods. The basic scheme uses a bag-of-features approach, where each image of the database is represented by a frequency vectors. The database images returned are those for which the associated frequency vectors are the k-nearest neighbors of the query frequency vector.

A major problem of this scheme is that the notion of neighborhood is not symmetric for the k-nearest neighbors search. Hence, if a vector x is a k-nearest neighbor of y, in general y is not a nearest neighbor of x.

Based on this observation, we have designed a contextual dissimilarity measure (CDM) which enhances the symmetry of the k-nearest neighbors relationship by taking into account the distance distribution of a given vector neighborhood. The performance of our approach is illustrated by Fig. 5. It shows some queries for which the CDM significantly improves the results.

Figure 5. Relevance of the proposed dissimilarity measure. For an object recognition benchmark (first two lines), the query with no CDM returns flowers, which are often irrelevantly returned. The query on the Lola database (two last lines) is even more impressive. The two first images are correct with and without CDM. Although the four next images seem wrong for both queries, they are in fact correct for the CDM, as the images correspond to the same location (the Deutsche Transfer Bank) observed from significantly different viewpoints.

6.1.6. Enriching local descriptors with color information

Local invariant descriptors are an efficient tool for scene representation due to their robustness with respect to occlusion and geometrical transformations. They are typically computed in two steps, detection of salient and sufficiently invariant local features, followed by the extraction of a robust visual appearance descriptor at the feature location and scale – see fig. 6. To be maximally informative, the descriptor should capture both the visual shape and the color characteristics of the local image region. A considerable amount of research has been dedicated to robust local shape descriptors, the SIFT descriptor [53] being the current reference. In our research we enrich the SIFT descriptor with a color description of the local region [50]. The color is represented by local histograms of photometric invariants. Furthermore, to counter the instabilities of nonlinear color transformations we use weights based on an error-analysis to robustify the histograms. An important subset of color descriptors is based on color derivatives. These descriptions have the disadvantage that they are dependent on image blur which can be caused by out-of-focus, relative movement between camera and object or rescaling of the image. In a multi-scale local feature setup, as presented in Fig. 6, features are rescaled to a standard size before the description phase. This rescaling introduces blur, making it imperative for its descriptions to be robust to image blur. We therefore derived a set of photometric invariant blur robust descriptors [49].

6.1.7. Discriminative regions for semi-supervised object class localization

Salient regions defined by local interest points are not sufficient to detect and characterize objects. First, they are sparse and can therefore not label every pixel. Second, they do not capture region-based

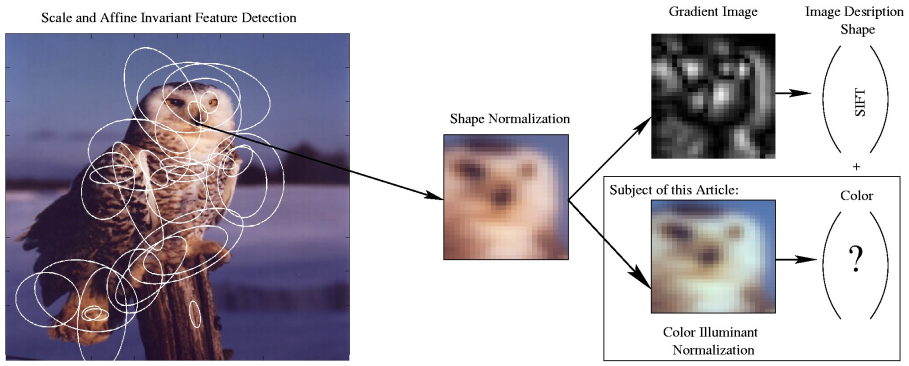


Figure 6. The local invariant descriptor method combines invariant local feature detection with robust local image description. The aim of this research is to enrich the local feature description with color information.

properties. Here we introduce a method which combines regions generated by image segmentation with local patches [42]. Region-based descriptors can model and match regular textures reliably, but fail on parts of the object which are textureless. They also cannot repeatably identify interest points on their boundaries. By incorporating information from patch-based descriptors near the regions into a new feature, the Region-based Context Feature (RCF), we can address these issues. We apply Region-based Context Features in a semi-supervised learning framework for object detection and localization. This framework produces object-background segmentation masks of deformable objects. Some examples of the results are given in fig. 7.

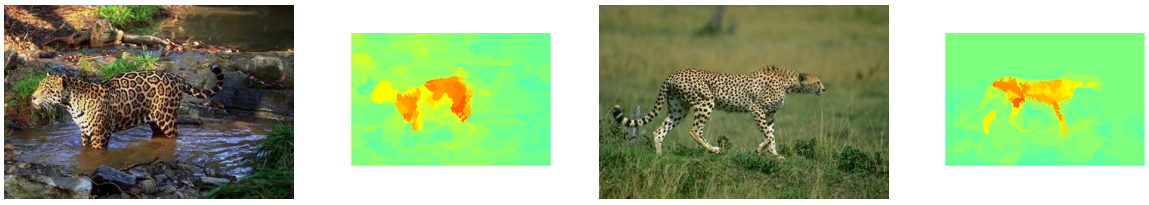


Figure 7. Examples of detections of spotted cats. The probability of a pixel being part of a spotted cat ranges from dark blue (low) through green (neutral) to dark red (high).

6.1.8. Maximally stable local description for scale selection

Recent work on image description has concentrated on improving invariance to geometric transformations by extracting invariant image regions and using these as support for descriptor calculation. The extraction and description steps are usually decoupled. However, it would be better to use a descriptor that is adapted to the detector. In fact, small changes in scale or location of the selected region can significantly alter the final descriptor. Scale selection turns out to be particularly sensitive. We have developed a detector that uses the descriptor itself to select the characteristic scales.

Our feature detector has two stages. An interest point detector is run at multiple scales to determine informative and repeatable locations. For each position we then apply a descriptor-based scale selection algorithm to identify maximally stable representations. The chosen scales are the ones at which the descriptor (here SIFT) *changes most slowly* as a function of scale. We call this algorithm *Maximally Stable Local Description* [31]. It performs well under changes of viewpoint and lighting conditions, often giving better

repeatability than other state-of-the-art methods. In our tests on object category classification is achieved similar or better results on four different datasets while for texture classification it always outperformed existing detectors. Stable orientation estimation (based on the dominant gradient at the keypoint location) can also be integrated into the method and again improves the texture classification results.

6.2. Statistical modeling and machine learning for image analysis

Participants: Charles Bouveyron, Juliette Blanchet, Hakan Cevikalp, Matthieu Guillaumin, Diane Larlus, Eric Nowark, Jakob Verbeek, Frédéric Jurie, Cordelia Schmid, Bill Triggs, Laurent Zwald, Florence Forbes [MISTIS], Stéphane Girard [LCM], Juho Kannala [University of Oulu].

6.2.1. Scene understanding using spatial statistical models

Automatically segmenting an image of a scene into the appearing concepts (such as trees, sky, building, cars, road, etc.) is a powerful tool since it not only determines what is present in a scene but also where it appears. Existing approaches are mainly based on estimating statistical model on the basis of a collection of manually segmented images. Clearly, manually producing a segmentation is a time consuming task. We have shown that using statistical aspect models (which were originally developed for text analysis) it is possible to estimate models for scene segmentation on the basis of a collection of weakly labelled images. For the weakly labelled images it is only indicated which concepts they contain –and not where in the image– and they thus require far less manual effort to generate, but contain significantly less information.

Aspect models regard the content of an image as drawn from a distribution which is a weighted sum of the distributions associated with the different concepts. The weak image labels used to estimate the models are leveraged by setting the weights of concepts not present in the image to zero. Model estimation iterates two steps until (guaranteed) convergence in an Expectation Maximization algorithm. In the first step the image is segmented probabilistically using all possible concepts according to the weak labeling. Figure 8 shows an illustration of an image, labelled building, grass, tree, sky, during model estimation (each pixel is weighted by its probability that it belongs to a certain concept). In the second step for each concept the associated distributions over image features calculated in the image (SIFT to describe the local gradient structure, and a robust hue measurement to describe local color content) are re-estimated using the probabilistic segmentation of the first step.

Figure 8. An example image used for model estimation labeled building, grass, sky, tree, together with its probabilistic segmentation over the four labels used during model estimation.

Although aspect models are useful, they ignore the spatial organization of the image as they model all measurements from the image as independent given the proportions of the image covered by the different concepts. By including spatial dependencies in the aspect model by means of a Markov random field structure, which renders the concepts displayed in nearby image regions dependent, we were able to further increase the segmentation performance. Notably, models including spatial dependencies estimated from weakly labelled images yield better segmentation results than a normal aspect model estimated from manually segmented images. This shows that by using a statistical model which is more appropriate for the image domain, the demands on the data required for the parameter estimation can be substantially reduced. Figure 9 shows several images and their segmentation produced by the spatial aspect model.

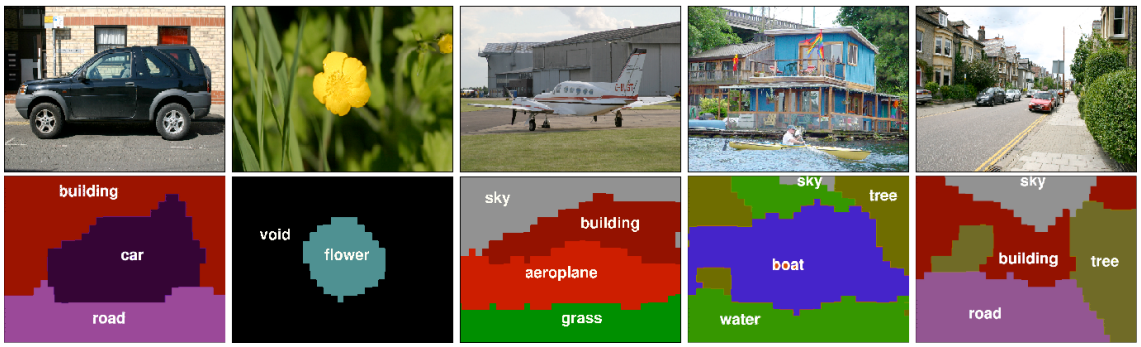


Figure 9. Examples of images segmented using a model for 20 different labels, each pixel is colored based on its most likely label.

6.2.2. Randomized clustering forests for building fast and discriminative visual vocabularies

Some of the most effective recent methods for content-based image classification work by extracting dense or sparse local image descriptors, quantizing them according to a coding rule such as k-means vector quantization, accumulating histograms of the resulting "visual word" codes over the image, and classifying these with a conventional classifier such as a SVM. Large numbers of descriptors and large codebooks are needed for good results and this becomes slow using k-means.

There are various methods for creating visual codebooks. K-means clustering is currently the most common but mean-shift and hierarchical k-means clustering have some advantages. These methods are generative but some recent approaches focus on building more discriminative codebooks. The above methods give excellent results but they are computationally expensive due to the cost of assigning visual descriptors to visual words during training and use. Tree based coders are quicker but (so far) somewhat less discriminative. It seems to be difficult to achieve both speed and good discrimination.

Our work contributes two main ideas [39]. One is that (small) ensembles of trees eliminate many of the disadvantages of single tree based coders without losing the speed advantage of trees. The second is that classification trees contain a lot of valuable information about locality in descriptor space that is not apparent in the final class labels. One can exploit this by training them for classification then ignoring the class labels and using them as "clustering trees" – simple spatial partitioners that assign a distinct visual word to each leaf.

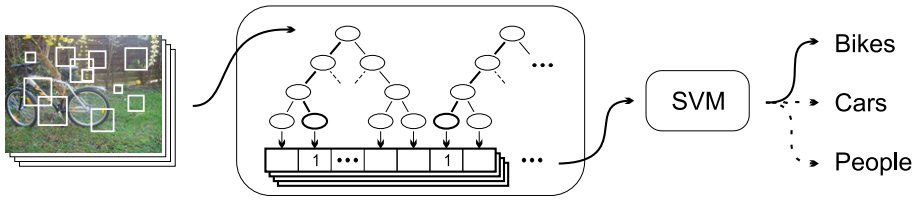


Figure 10. Randomized clustering forests scheme.

Combining these ideas, we introduce Extremely Randomized Clustering Forests (ERC-Forests) – ensembles of randomly created clustering trees. We show that these have good resistance to background clutter and that they provide much faster training and testing and more accurate results than conventional k-means in several state-of-the-art image classification tasks. The overall scheme is illustrated by Figure 10.

6.2.3. High dimensional data analysis and clustering

The visual descriptors used in object recognition are usually high-dimensional, but often lie in different low-dimensional subspaces of the original space. Global dimensionality reduction techniques are not useful in this case. We propose a method for discriminant analysis and clustering that finds the specific subspace and intrinsic dimension of each class. Our approach adapts a mixture of Gaussian framework to high-dimensional data: It determines class-specific subspaces and therefore limits the number of parameters that need to be estimated. The intrinsic dimension of each class is determined automatically using Cattell's scree test on eigenvalue sizes. The resulting approach for discriminant analysis [6], [7] has shown good results for classification of visual descriptors, i.e., it outperforms linear SVMs (support vector machines).

The extension to clustering uses EM for parameter estimation and provides a robust clustering method in high-dimensional spaces that we call High Dimensional Data Clustering (HDDC) [28]. We applied the method to probabilistic object recognition. Local scale-invariant features are clustered using HDDC, and the maximum likelihood based discriminative score for each cluster is determined using positive and negative examples of the category. This allows to assign object probabilities to each visual descriptor. Localization then assumes that points with higher probabilities are more likely to belong to the object, and image classification is based on a per image score of the probabilities. Experiments on two recent object databases demonstrate the effectiveness of the clustering method for category-level localization and classification [29], see figure 11 for example results.

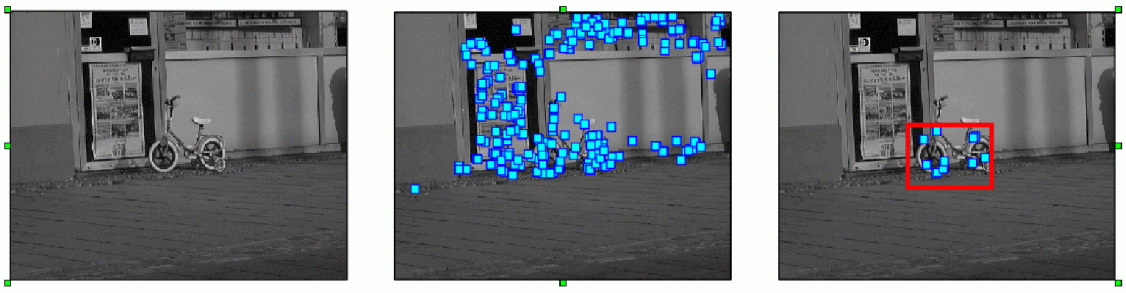


Figure 11. Left: Image. Middle: Detected interest points. Right: Most discriminative points and object localization result.

6.2.4. Latent mixture vocabularies for object classification

The visual vocabulary is an intermediate level representation which has been proved to be very powerful for addressing object categorization problems. It is generally built by vector quantizing a set of local image descriptors during a distinct preprocessing stage without taking into account the model. However, image representation and thus classification performance strongly depend on it. The efficiency of vocabularies estimated without consideration of either the classification task nor the image modeling process is not optimal. We, therefore, proposed a generative model to describe images where the visual vocabulary is a built-in component of the model, learned simultaneously with other parameters [35]. The model is based on latent aspects (more precisely on the LDA model). Some hidden variables (the topics or aspects) are responsible for generating all the observations, which are the visual words in our case. The visual words are modeled as Gaussian mixtures of visual descriptors.

This model is estimated on a set of images in order to produce for each image a distribution over the topics and for each visual word the corresponding Gaussian parameters. Some supervision is introduced in the model estimation to make the topics depend on the classes. These topics and visual words are components of the model: on one hand the topics (related to classes) will influence the creation of words and make them more

suited for the classification task; on the other hand the visual words will be adapted to the model and make the topic estimation more precise. Experimental results show that our approach outperforms the ones where visual words are learnt without during a separate preprocessing stage.

6.2.5. Learning visual distance

We introduced an algorithm that learns a similarity measure for comparing objects which haven't been observed before [40]. The problem addressed is illustrated by Figure 12. The measure is learned from pairs of training images labeled "same" or "different". This is far less informative than fully labeled images (e.g. "car model X") used in general, but cheaper to obtain. The computation of the similarity measure is a two folds process in which local visual descriptors from both images are vector quantized by an ensemble of extremely randomized binary trees and compared using a classifier. These trees are very fast to learn and are also robust because of the redundant information they embed. Furthermore, they automatically select and combine descriptor features (gray-scale pixel information, gradient, geometry, etc) most adapted to each case. We evaluated our similarity measure on four different datasets and always outperformed the state-of-the-art results.

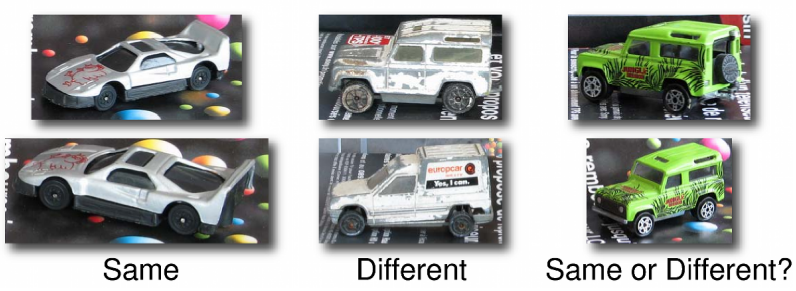


Figure 12. Given pairs labeled "same" or "different", we learn a similarity measure that decides if two images represent the same object. The similarity measure should be robust to modifications in pose, background and lighting conditions, and should deal with never seen objects.

6.2.6. Local subspace classifiers: linear and nonlinear approaches

The K-local hyperplane distance nearest neighbor algorithm (HKNN) is a successful local classification method which builds nonlinear decision surfaces directly in the original sample space rather than in the feature space. It is not possible to employ distance metrics other than the Euclidean distances in this scheme, which is a major limitation of the method. We re-formulated the HKNN method in terms of subspaces. The advantages of such a subspace formulation are two-fold: First, it enables us to propose a variant of the HKNN algorithm, the Local Discriminative Common Vector (LDCV) method, which is more suitable for classification tasks where classes have similar intra-class variations. Second, the HKNN method can be extended to the nonlinear case based on subspace concepts. However, the non-linearization of the method was not trivial. The subspace method needed to be modified before the non-linearization. As a result of the non-linearization process, one may utilize a wide variety of distance functions in those local classifiers. We tested the proposed methods, namely the nonlinear HKNN (NHKNN) and the nonlinear LDCV (NLDCV), on several classification tasks. Experimental results showed that the proposed methods yield comparable or better results than the Support Vector Machine (SVM) classifier and its local counterpart SVMKNN.

We also proposed a new supervised clustering algorithm, named the Homogeneous Clustering (HC), which finds the number and initial locations of the hidden units in the Radial Basis Function (RBF) neural network classifiers. The experimental results showed that the RBF network classifier performs better when it is initialized with the proposed HC algorithm rather than an unsupervised k-means algorithm.

6.3. Visual object recognition

Participants: Ankur Agarwal, Navneet Dalal, Diane Larlus, Marcin Marszalek, Vittorio Ferrari, Alexander Kläser, Xiaoyang Tan, Joost Van de Weijer, Frédéric Jurie, Roger Mohr, Cordelia Schmid, Bill Triggs, Akash Kushal [UIUC], Svetlana Lazebnik [UIUC], Jean Ponce [UIUC & ENS Ulm].

6.3.1. Image classification with bag-of-features

Many current approaches for image classification are based on visual words (clusters of local descriptors) frequency vectors and SVM classifiers. They obtain excellent results, see for example Figure 13 for classification results for unknown test images.

We have made a large-scale evaluation of the sparse approach [22]. It represents images as distributions (signatures or histograms) of features extracted from a sparse set of keypoint locations and learns a Support Vector Machine classifier with kernels based on two effective measures for comparing distributions, the Earth Mover's Distance and the χ^2 distance. A comparative evaluation with several state-of-the-art recognition methods on four texture and five object databases showed that our implementation exceeds the best reported results on most of them and achieves comparable performance on the rest. Our experiments demonstrate that image representations based on distributions of local features are surprisingly effective for classification of texture and object images under challenging real-world conditions, including significant intra-class variations and substantial background clutter.

More recently we have introduced several improvements [35], [39], [41]. First, we extract local descriptors based on multi-scale dense sampling. Such a representation is more complete, hence more representative of the image content. Second, we obtain more significant visual words based on learning techniques which take into account the class membership of the descriptors.

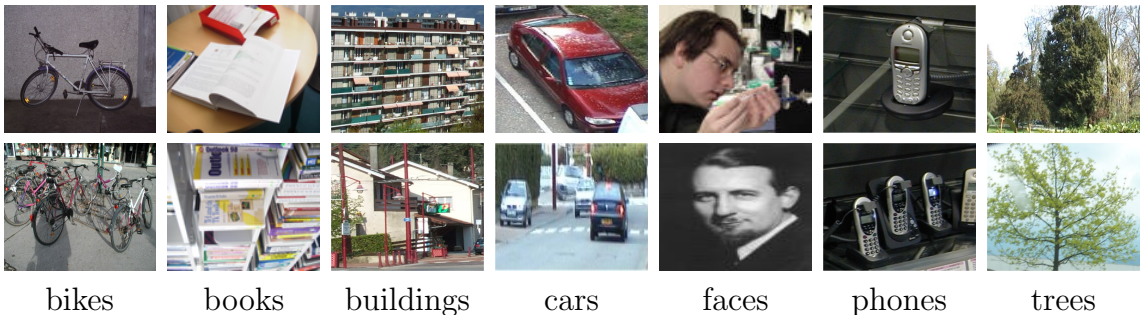


Figure 13. Images of categories bikes, books, buildings, cars, faces, phones and trees of the Xerox7 dataset. Note that all of them are classified correctly with one of our approaches [22].

6.3.2. Spatial Pyramid Matching

Standard bag-of-features approaches do not take into account the spatial information of the image descriptors, which is discriminant. We have, therefore, developed an approach which includes the spatial layout based on approximate global geometric correspondence [36]. This technique works by partitioning the image into increasingly fine sub-regions and computing histograms of local features found inside each sub-region, see figure 14(a). The resulting "spatial pyramid" is a simple and computationally efficient extension of an orderless bag-of-features image representation, and it shows significantly improved performance on challenging scene categorization tasks, as illustrated by Figure 14(b). Specifically, our proposed method exceeds the state of the art on the Caltech-101 database and achieves high accuracy on a large database of fifteen natural scene categories. The spatial pyramid framework also offers insights into the success of several recently proposed

image descriptions, including Torralba's "gist" and Lowe's SIFT descriptors. This is joint work with S. Lazebnik (UIUC) and J. Ponce (UIUC and ENS Ulm).

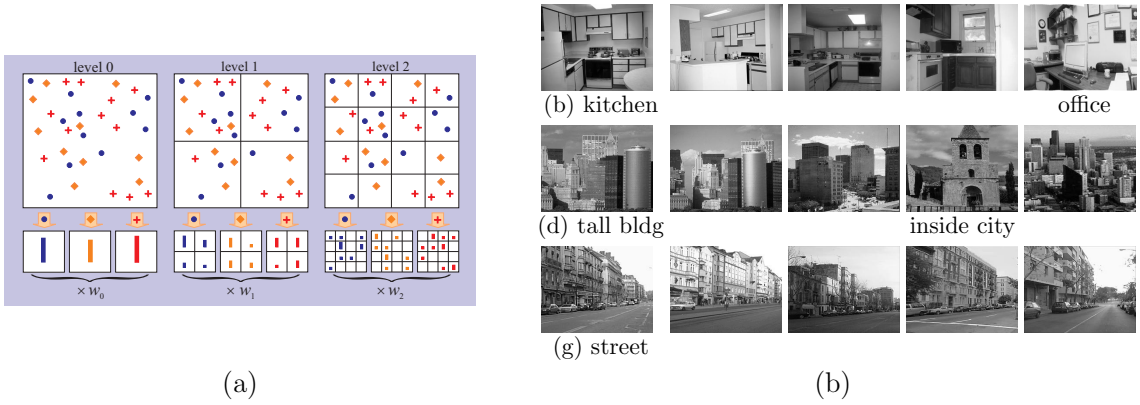


Figure 14. (a) Toy example of constructing a three-level pyramid. The image has three feature types, indicated by circles, diamonds, and crosses. At the top, we subdivide the image at three different levels of resolution. Next, for each level of resolution and each channel, we count the features that fall in each spatial bin. Finally, we weight each spatial histogram. (b) Retrieval from the scene category database. The query images are on the left, and the four images giving the highest values of the spatial pyramid kernel are on the right. The actual class of incorrectly retrieved images is listed below them.

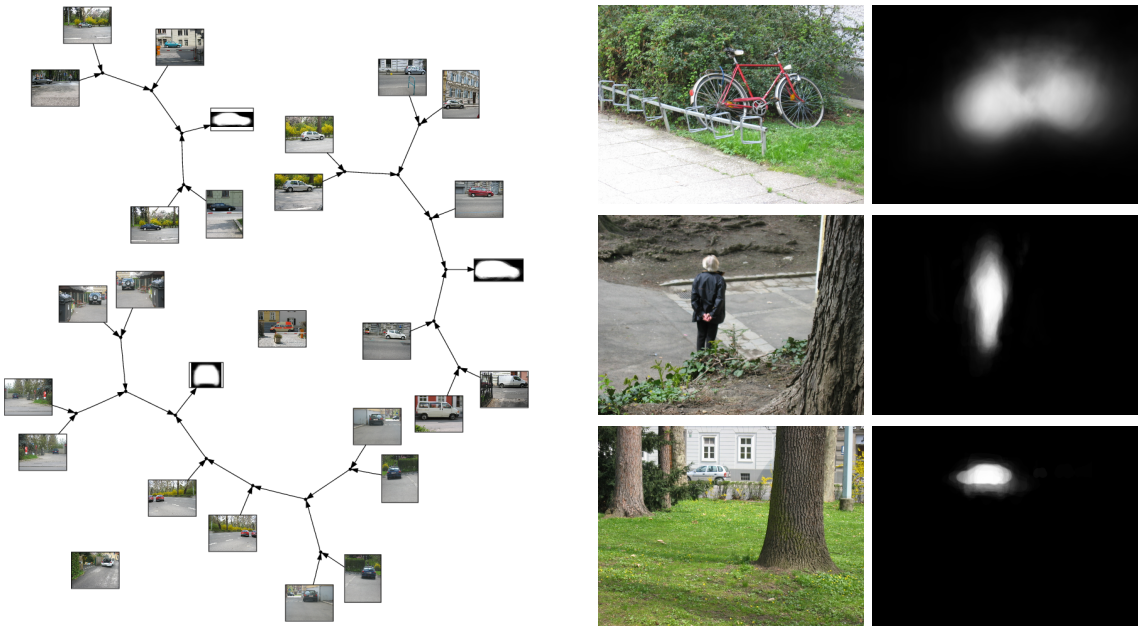
6.3.3. Spatial weighting for bag-of-features and localization

In the following we present another approach which adds spatial relations to a bag-of-features representation [37]. The idea is to exploit spatial relations between features based on object boundaries provided during supervised training. Each feature of a test image votes for the aligned object shapes of the corresponding features in the training images. We boost the weights of features which agree on the position and shape of the object and suppress the weights of background features, hence the name of our method—"spatial weighting". The proposed representation is thus richer and more robust to background clutter. Experimental results show that "spatial weighting" improves the results when added to a bag-of-features classification technique.

Furthermore, it is possible to apply the "spatial weighting" to object localization. Our approach goes beyond bounding boxes, as it determines the outline of the object. It also learns and detects possible object viewpoints and articulations, which are often well characterized by the object outline. Unlike most current localization methods, we do not require any hypothesis parameter space to be defined. Instead, our approach directly generates, evaluates and clusters shape masks. To reduce the number of training shapes, we perform clustering of the individual object shapes, see Figure 15(a). We evaluated the proposed approach on the challenging natural-scene Graz-02 object classes dataset. The results demonstrate the extended localization capabilities of our method, see Figure 15(b).

6.3.4. Category-level object segmentation

In the following, we present an approach for segmenting objects of a given category, where the category is defined by a set of segmented training images—a problem commonly refered to as the *figure-ground segmentation*. Our method represents images and objects with a latent variable model similar to the Latent Dirichlet Allocation (LDA) model. We extended LDA by defining images as multiple overlapping regions, each of which is considered as a distinct document. This gives a higher chance to small objects of being



(a) (b)

Figure 15. (a) Main car aspects detected by agglomerative clustering. (b) Example localization results on the Graz-02 dataset.

discovered, i.e., they are more likely to be the main topic of an image sub-region. This overlapping scheme also enforces cooperation between documents and leads to a better estimation of the class for each image patch due to partial coherency. This model is well-suited for assigning image patches to objects (even if they are small), and therefore for segmenting objects. Indeed, each pixel can be assigned to a class by averaging information from all patches it belongs to. We then obtain pixel-wise probability maps for foreground and background. Putting a threshold on this probability map generates accurate segmentation masks, as shown in Figure 16.

Figure 16. Binary segmentation masks for the bike category produced by the proposed method, see text.

6.3.5. Hyperfeatures – multilevel local coding for visual recognition

Visual words are not able to exploit spatial co-occurrence statistics at scales larger than their local input patches. We have developed a new multilevel visual representation, 'hyperfeatures', that is designed to remedy this [26]. The starting point is the familiar notion that to detect object parts, in practice it often suffices to detect co-occurrences of more local object fragments – a process that can be formalized as comparison (*e.g.* vector quantization) of image patches against a codebook of known fragments, followed by local aggregation of the resulting codebook membership vectors to detect co-occurrences. This process converts local collections of image descriptor vectors into slightly less local histogram vectors – higher-level but spatially coarser descriptors. As the output is again a local descriptor vector, the process can be iterated, and doing so captures and codes ever larger assemblies of object parts and increasingly abstract or 'semantic' image properties. We have studied the performance of hyperfeatures extensions of several different image coding methods including clustering based Vector Quantization and Gaussian Mixtures, finding that the latter consistently outperform the former, and that adding a stage of LDA – a probabilistic "topic distillation" model that has recently been developed in the statistical text community – improves the results further. The resulting high-level features provide improved performance in several object image and texture image classification tasks. We are currently in the process of developing the method for object localization. Some results demonstrating this are shown in Figure 17.

6.3.6. Accurate object detection with deformable shape models learnt from images

Most recent approaches to object detection localize objects up to a rectangular bounding-box. Here, we want to go a step further and localize object *boundaries*. Our approach bridges the gap between shape matching and object detection. Classic non-rigid shape matchers obtain accurate point correspondences, but take *point sets* as input. In contrast, we build a shape matcher with the input/output behavior of a modern object detector: it learns shape models *from images*, and automatically localizes them in cluttered images. This is possible due to (i) a novel technique for learning a shape model of an object class given *images* of example instances; (ii) the combination of Hough-style voting with a non-rigid point matching algorithm to localize the model

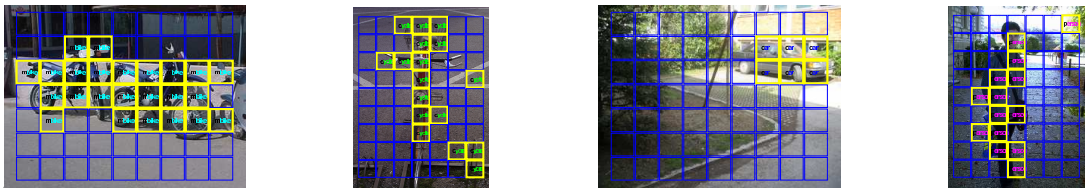


Figure 17. Object localization on 4 object categories from the PASCAL object recognition challenge dataset, based on classifying each local image region using its hyperfeatures. Currently, each region is classified independently – no spatial smoothness constraint is enforced.

in cluttered images. As demonstrated by an extensive evaluation, our method can localize object boundaries accurately. Training does not require segmented examples (only bounding-boxes). Figure 18 (top row) shows a few localization results. We can observe the objects are localized very accurately. The bottom row displays two models instances for each class. Each model instance is learnt from a different set of training images. Note that the models are robust to variations in the training images.

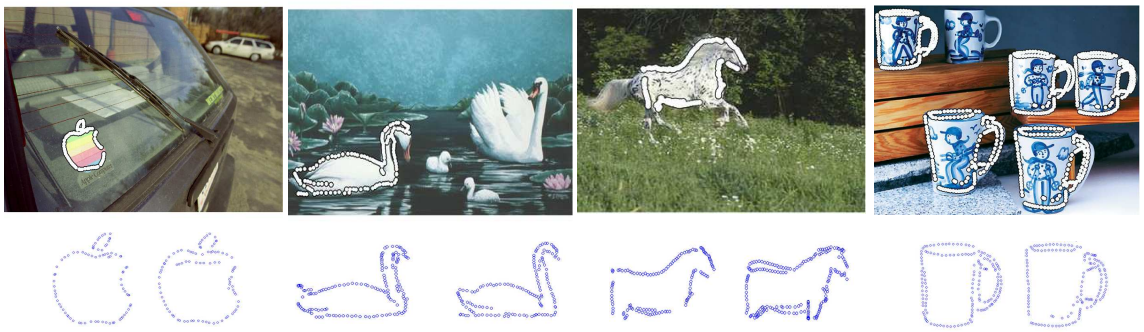


Figure 18. The top row shows a few localization results. Note the accuracy of the detected outlines. The bottom row shows two models for each class—they are learnt from different set of training images.

6.3.7. Flexible object models for category-level 3D object recognition

Today's category-level object recognition systems largely focus on centered fronto-parallel views of nearly rigid objects with characteristic texture patterns. To overcome these limitations, we propose a novel framework for visual object recognition where object classes are represented by graphs of *partial surface models* (PSMs) obeying loose local geometric constraints, see figure 19 for an illustration. Our model only enforces *local* geometric consistency, both at the level of model parts and at the level of individual features within the parts, and it is therefore robust to viewpoint changes and intra-class variability.

The PSMs themselves are formed of dense, locally rigid assemblies of image features. They are learned by matching repeating patterns of features across training images of each object class. Pairs of PSMs which regularly occur near each other at consistent relative positions are then linked by edges whose labels reflect the local geometric relationships between these features. These local connections are used to construct a probabilistic graphical model for the geometry and appearance of the PSMs making up an object. The corresponding *PSM graph* is the basis for an effective algorithm for object detection and localization, which

outperforms the state-of-the-art methods on the Pascal 2005 VOC challenge cars test 1 data. This is joint work with A. Kushal (UIUC) and J. Ponce (UIUC and ENS Ulm).

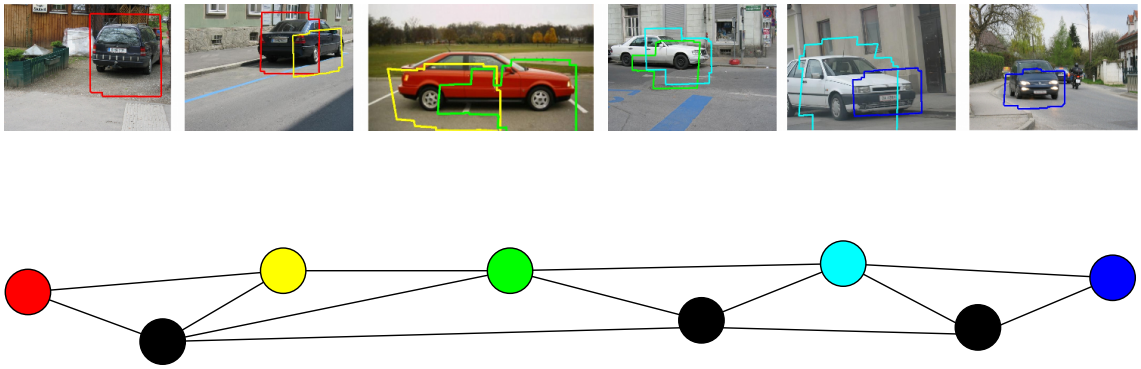


Figure 19. An example of a learned PSM graph. The top row shows the outlines of the PSM instances corresponding to nodes with the same color in the PSM graph below it. The black nodes represent other nodes in the PSM graph.

6.3.8. Learning color names from real world images

Our current research includes research on automatic learning of color names. Color names are linguistic labels that humans attach to colors. The current paradigm to obtain color names is to label a collection of color chips with color names within a well-defined experimental setup. In contrast we propose to learn the color names automatically from images retrieved by Google image. To learn the color red, we query Google image for "red+color". This process is repeated for a fixed set of colors after which we apply Probabilistic Latent Semantic Analysis (PLSA) to learn the color names. The learned color names are tested for the retrieval of objects, e.g., retrieve all "red cars", and compared to a 'traditional' set of color names learned within a laboratory setup. Results show that the color names learned with our method outperform color names learned in a laboratory setup. This is probably due to the fact that color-naming within a laboratory setup of individual color chips does not resembles color naming in real-world images. On the other hand the color names learned by our method are well-fit to the task of color naming in the real-world.

6.3.9. Human detection based on histogram of oriented gradient descriptors

We have developed a robust feature set for visual object recognition in general, and for "pedestrian detection" (the detection of upright, full visible, standing or walking people) in particular [4]. Our "Histogram of Oriented Gradient (HOG)" feature set is inspired by the success of SIFT descriptors [53] for local feature based recognition, but here, rather than being sampled only sparsely at salient local feature points, blocks of well-normalized gradient orientation histograms are used in a dense grid, at a uniform scale and without rotation normalization. This provides a particularly robust and discriminant appearance based descriptor suitable for visual classes that have a reasonable degree of spatial consistency across examples. The current overall detector uses a monolithic (non-parts-based) linear Support Vector Machine as a classifier in the HOG window, scanning this densely across the test image at multiple scales followed by a mean-shift based space-scale peak location method to recover multi-scale detections. We have extended the method to include a differential optical flow channel to provide improved discrimination for human detection in films and videos [30]. The method allows stationary or moving subjects, non-rigid or moving backgrounds and moving cameras. It provides a further order of magnitude reduction in false positive rate.

7. Contracts and Grants with Industry

7.1. Bertin Technologies

Participants: Eric Nowak, Frédéric Jurie, Roger Mohr.

The collaboration with Bertin Technologies focuses on developing algorithms for detecting and recognizing objects in unmanned infra-red information systems. Typical applications are outdoors defense systems in which hidden cameras are left to detect the presence of military vehicles. The main challenges are the relatively poor image resolution, the changeable appearance of objects due to global and local temperature changes, and the potentially large number of nested object categories. The project funds the CIFRE grant for Eric Nowak's PhD thesis, which started in March 2004. Bertin Technologies also participates in our Techno-Vision project ROBIN (see paragraph 8.1.2).

7.2. MBDA Aerospatiale

Participants: Julien Bohne, Frédéric Jurie, Cordelia Schmid.

We have collaborated with the Aerospatiale section of MBDA for several years. In November 2005, we started a one year transfer contract. During 2006, we studied three issues for infra-red images: registration under large view point changes, the evaluation of keypoint based detection and matching, and the tracking of small objects. In December 2006 we started a three-year contract on object localization and classification. MBDA also participates in our Techno-Vision project ROBIN.

7.3. EADS Fondation

Participants: Vittorio Ferrari, Frédéric Jurie, Cordelia Schmid.

The postdoctoral scholarship of Vittorio Ferrari was financed by the EADS Foundation. The project started in November 2005 and explored image contours as an alternative representation for visual class recognition. In contrast to systems based on local invariant textured patches, this allows to recognize classes that are mostly defined by their shape, such as bottles, mugs, or horses. Furthermore, we developed a method which automatically learns shape models from images.

8. Other Grants and Activities

8.1. National Projects

8.1.1. Ministry grant MoViStaR

Participants: Charles Bouveyron, Juliette Blanchet, Cordelia Schmid.

MoViStaR is a joint national project ("action concertée incitative") under the "Masses de Données" (Processing Large Datasets) program. The partners are LEAR (C. Schmid), INRIA's MISTIS team (F. Forbes), the SMS team of the LMC laboratory (S. Girard) and the Heudiasyc laboratory (C. Ambroise). The project started in September 2003 for three years. It aimed at developing techniques to achieve reliable category-level visual recognition by mining information from large image collections. Particular focuses are developing and adapting advanced statistical data reduction techniques and integrating spatial information into the process. C. Bouveyron who was financed by the project successfully defended his Ph.D on high dimensional data analysis and clustering in September 2006.

8.1.2. Techno-Vision project ROBIN

Participants: Benjamin Ninassi, Hakan Cevikalp, Frédéric Jurie, Roger Mohr.

We lead the national Techno-Vision project ROBIN, which started in January 2005 for two and a half years. Its aim is to quantify and consolidate progress in visual object recognition by developing ground truth

datasets and performance metrics to improve the evaluation of object recognition algorithms, and by running a national competition in this area. The project is funded partly by the French Ministry of Defense, the French Ministry of Research and by several companies and research centers (Bertin Technologies, Cybernetix, DGA, EADS, INRIA, ONERA, MBDA, SAGEM, THALES and 35 public laboratories). It covers multi-class object detection, generic object detection, generic object recognition, and image categorization. During the first year the project produced datasets and metadata, while the second year was devoted to selecting the test images and the benchmarking procedure as well as organizing the competition. The actually competition will take place in 2007.

8.2. European Projects and Grants

8.2.1. FP6 Integrated Project aceMedia

Participants: Navneet Dalal, Matthijs Douze, Bill Triggs, Cordelia Schmid.

AceMedia is a 6th framework integrated project that is running for 4 years starting from January 2004. It aims to integrate knowledge, semantics and content for user-centered intelligent media services. The partners are: Motorola Ltd UK (coordinator); Philips Electronics Netherlands; Thomson France; Queen Mary College, University of London; Fraunhofer FIT; Universidad Autonóma de Madrid; Fratelli Alinari; Telefónica Investigación y Desarrollo; the Informatics and Telematics Institute, Dublin City University; INRIA (including the TexMex team at IRISA in Rennes, Imedia at Rocquencourt in Paris, and LEAR in Grenoble); France Télécom; Belgavox; the University of Karlsruhe; Motorola SAS France. LEAR has worked on human detection and action recognition in static images and in videos. In 2006 it started a second branch of work on the semi-automatic organization of home photo collections.

8.2.2. FP6 Project CLASS

Participants: Yves Gufflet, Alexander Kläser, Xiaoyang Tan, Jakob Verbeek, Cordelia Schmid, Bill Triggs.

CLASS (Cognitive-Level Annotation using latent Statistical Structure) is a 6th framework Cognitive Systems STREP that started in January 2006 for three years, coordinated by LEAR. It is a basic research project focused on developing a specific cognitive ability for use in intelligent content analysis: the automatic discovery of content categories and attributes from unstructured content streams. It studies both fully autonomous and semi-supervised methods. The work combines robust computer vision based image descriptors, machine learning based latent structure models, and advanced textual summarization techniques. The potential applications of the basic research results are illustrated by three demonstrators: an Image Interrogator that interactively answers simple user-defined queries about image content; an automatic annotator for people and actions in situation comedy videos; an automatic news story summarizer. The Class consortium is interdisciplinary, combining five leading European research teams in visual recognition, text understanding and summarization, and machine learning: LEAR; Oxford University, UK; K.U. Leuven, Belgium; T.U. Darmstadt and MPI Tuebingen, Germany.

8.2.3. FP6 Network of Excellence PASCAL

Participants: Ankur Agarwal, Juliette Blanchet, Charles Bouveyron, Navneet Dalal, Marcin Marszałek, Frédéric Jurie, Cordelia Schmid, Bill Triggs.

PASCAL (Pattern Analysis, Statistical Modeling and Computational Learning) is a 6th framework EU Network of Excellence that started in December 2003 for four years, funded initially by the European Commission's Multimodal Interfaces unit, and currently by its Cognitive Systems one. The focus is on applying advanced machine learning and statistical pattern recognition techniques to the analysis of various types of sensed data. It currently unites about 540 researchers, postdocs and students from 56 sites, mainly in Europe but also including sites in Israel and Australia. Subject areas covered include machine learning, statistical modeling and pattern recognition, and application domains including computer vision, natural language processing including speech, text and web analysis, information extraction, haptics and brain computer interfaces. The coordinator is John Shawe-Taylor of Southampton University. Bill Triggs coordinates

the computer vision aspects and manages various activities including the Balance & Integration and Funding Review Programs. LEAR and Xerox Research Center Europe (XRCE) together form one of PASCAL's 14 key sites, focusing on computer vision and natural language processing.

8.2.4. FP6 Marie Curie EST host grant VISITOR

Participants: Marcin Marszałek, Cordelia Schmid.

LEAR is one of the teams participating in VISITOR, a 3 year Marie Curie Early Stage Training Host grant of the GRAVIR-IMAG laboratory to which LEAR belongs. VISITOR is funding the PhD of the Polish student Marcin Marszałek.

8.2.5. EU Marie Curie EST grant PHIOR

Participants: Joost Van de Weijer, Cordelia Schmid.

PHIOR is a Marie Curie postdoctoral grant for J. Van de Weijer on photometric robust features for object recognition in color images. It runs since November 2005 for two years. The project aims at improving image descriptors by adding robust color information. Machine learning techniques determine the most discriminative features, e.g. choosing between different levels of invariance and features types. Furthermore, we learn the colorimetric properties of images and categories.

8.3. Bilateral relationships

8.3.1. University of Illinois at Urbana-Champaign, USA

Participants: Cordelia Schmid, Bill Triggs, Akash Kushal [UIUC], Svetlana Lazebnik [UIUC], Jean Ponce [UIUC & ENS Ulm].

This collaboration on 3D object recognition between the research group of J. Ponce and LEAR is funded by a CNRS/INRIA/UIUC collaboration agreement. In 2006, C. Schmid visited the partner institution for one week and was part of S. Lazebnik's Ph.D. committee. Furthermore, she worked with A. Kushal during his visit to Paris. B. Triggs also visited UIUC for one week and J. Ponce came to Grenoble twice for one day.

9. Dissemination

9.1. Leadership within the scientific community

- Conference and workshop organization:
  - Co-organizer of Workshop on Category-level Object Recognition, Siracusa, Sicily, September 2006 (C. Schmid)
  - Co-organizer of Workshop on Mathematical Methods in Computer Vision, Banff, Alberta, Canada, October 2006 (B. Triggs)
- Editorial boards:
  - International Journal of Computer Vision (C. Schmid)
  - IEEE Transactions on Pattern Analysis and Machine Intelligence (B. Triggs)
  - Foundation and Trends in Computer Graphics and Vision (C. Schmid)
- Area chairs:
  - NIPS'06 (C. Schmid)

  - ECCV'06 (B. Triggs)
  - CVPR'06 (B. Triggs)
  - CVPR'07 (C. Schmid)

- Program committees:

  - NIPS'06 (B. Triggs, J. Verbeek)
  - ECCV'06 (A. Agarwal, V. Ferrari)
  - RFIA'06 (C. Schmid)
  - CVPR'06 (V. Ferrari, F. Jurie, C. Schmid, J. Verbeek)
  - ICIP'06 (J. Van de Weijer)
  - CVPR'07 (H. Jegou, F. Jurie, B. Triggs, J. Van de Weijer, J. Verbeek)
  - VISAPP'07 (F. Jurie)
  - Acivs'06 (F. Jurie)

- Prizes:

  - Cordelia Schmid and Roger Mohr received the *Longuet-Higgins Prize for Fundamental Contributions in Computer Vision that Have Withstood the Test of Time*, June 2006.
  - Hervé Jégou received the national prize of the best PhD thesis in image and signal processing of the Club EEA (French Electrical Engineering Association).
  - Jakob Verbeek received the biannual E.S. Gelsema award of the Dutch Society for Pattern Recognition and Image Processing for best PhD thesis and associated international journal publications.
  - The paper "Object Localization by subspace clustering of local descriptors" by C. Bouveyron, J. Kannala, C. Schmid and S. Girard won the honorary mention prize at the Indian Conference on Computer Vision, Graphics and Image Processing.
  - Methods submitted by LEAR won 12 of the 20 categories of the Visual Object Recognition Challenge proposed by the European Network of Excellence PASCAL in April 2006.
  - LEAR participated in the ImagEVAL competition financed by the French "Techno-Vision" program in the task "recognition of transformed images" in October 2006. We came second for subtask one and first (tied) for subtask two.

- Other:

  - F. Jurie is vice-head of AFRIF (the French section of the IAPR).
  - F. Jurie is scientific co-director of GDR ISIS (the national interest group on image analysis).
  - C. Schmid is a member of INRIA's Commission d'Évaluation, and of the INRIA Rhône-Alpes Comité des Emplois Scientifiques. She has participated in several recruitment committees.
  - C. Schmid is in charge of international relations at INRIA Rhône-Alpes.
  - B. Triggs is a member of INRIA's COST (Scientific and Technical Strategy Committee).
  - B. Triggs is deputy director of the Laboratoire Jean Kuntzmann, a CNRS-INPG-UJF laboratory on applied mathematics and computer science, formed in January 2007.
  - B. Triggs manages the Funding Review and the Balance & Integration programmes of the EU Network of Excellence PASCAL, and co-manages several other programmes.

9.2. Teaching

- H. Jégou, Pattern Recognition, University de Rennes I, Master 2, 12h
- F. Jurie, Matching and Recognition, INPG, Masters IVR, 12h
- F. Jurie, Multi-media databases, INPG, 3rd year ENSIMAG, 18h
- D. Larlus, INPG Bachelor 1, Practical courses in functional programming, 40h
- D. Larlus, INPG, Mathematics for computer sciences, Master ICA, 2nd year, 22h

9.3. Invited presentations

- V. Ferrari, *Object Detection with Contour Segment Networks*, KU Leuven, VISICS group, April 2006
- V. Ferrari, *Object Detection with Contour Segment Networks*, University of Oxford, VGG group, May 2006
- H. Jégou, *Nearest neighbors search*, Reykjavik, Iceland, April 2006
- H. Jégou, *Improving projection-based approximate nearest neighbors search*, Workshop on Category-Level Object Recognition, Poster, Siracusa, September 2006
- F. Jurie, *Fast Discriminative visual codebooks using randomized clustering forests*, Workshop on Category-Level Object Recognition, Siracusa, September 2006
- F. Jurie, *Catégorisation d'images : avancées récentes*, Journee du GdR ISIS, Paris, October 2006
- D. Larlus, *Utilisation de modèles à variables latentes pour l'amélioration du vocabulaire visuel dans le cadre de la catégorisation d'objets*, Journee du GdR ISIS, Paris, October 2006
- M. Marszałek, *Bag-of-features and beyond*, Journee du LJK, Saint Quentin, September 2006
- M. Marszałek, *Spatial Weighting for bag-of-features*, Workshop on Category-Level Object Recognition, Poster, Siracusa, September 2006
- M. Marszałek, *Bag-of-features image representation: state-of-the-art and beyond*, Journee du GdR ISIS, Paris, October 2006
- M. Marszałek, *MoviStar: bag-of-features and beyond*, PaRISTIC, Nancy, November 2006
- R. Mohr, *Aller-retours entre théorie et réalité, ou une leçon de modestie en vision par ordinateur*, conférence invitée pour les 20 ans du Loria, December 2006
- E. Nowak, *Reconnaissance d'objets dans des images: une approche par sac-de-mots*, Journee du GdR ISIS, March 2006
- C. Schmid, *Bag-of-features and beyond*, Seminar at UIUC, Champaign, February 2006
- C. Schmid, *Bag-of-features and beyond*, Seminar at ETH Zurich, April 2006
- C. Schmid, *Bag-of-features and beyond*, Seminar at Xerox, Grenoble, June 2006
- C. Schmid, *Invariant local features*, Tutorial at AERFAI Summer School on Action and Object Classification Techniques in Digital Images, Granada, Spain, June 2006
- C. Schmid, *Beyond bag-of-features: adding spatial and shape information*, Seminar at Microsoft Research, Seattle, August 2006
- C. Schmid, *Beyond bag-of-features: adding spatial and shape information*, Seminar at Carnegie Mellon University, Pittsburg, August 2006
- C. Schmid, *Groups of adjacent contour segments for object detection*, Workshop on Category-Level Object Recognition, Siracusa, September 2006

- C. Schmid, *Beyond bag-of-features: adding spatial and shape information*, Seminar at ENS Ulm, October 2006
- B. Triggs, *Histograms of oriented gradients for human detection*, UIUC Champaign, March 2006
- B. Triggs, *Contributions to visual recognition and human motion estimation*, Australian National University, Canberra, August 2006
- B. Triggs, *Finding people in images and videos*, Seminar at Microsoft Research, Seattle, October 2006
- J. Van de Weijer, *Coloring local feature extraction for object recognition*, Computer Vision Center Barcelona, April, 2006
- J. Verbeek, *Unsupervised learning of low-dimensional structure in high-dimensional data*, Learning Workshop at the Laboratoire Lorrain de Recherche en Informatique, Nancy, February 2006
- Presentation of LEAR's image search demonstrator at the 2006 "Fête de la Science", Grenoble and at the Forum 4i, Grenoble, 2006

10. Bibliography

Year Publications

Books and Monographs

[1] J. PONCE, M. HEBERT, C. SCHMID, A. ZISSERMAN. *Towards category-level object recognition*, to appear, Springer, http://lear.inrialpes.fr/pubs/2006/PHSZ06.

Doctoral dissertations and Habilitation theses

[2] A. AGARWAL. *Machine learning for image based motion capture*, Ph. D. Thesis, Institut National Polytechnique de Grenoble, apr 2006, http://lear.inrialpes.fr/pubs/2006/Aga06a.

[3] C. BOUVEYRON. *Modélisation et classification des données de grande dimension : application à l'analyse d'images*, Ph. D. Thesis, Université Joseph Fourier, Grenoble 1, sep 2006, http://lear.inrialpes.fr/pubs/2006/Bou06.

[4] N. DALAL. *Finding people in images and videos*, Ph. D. Thesis, Institut National Polytechnique de Grenoble, july 2006, http://lear.inrialpes.fr/pubs/2006/Dal06.

Articles in refereed journals and book chapters

[5] A. AGARWAL, B. TRIGGS. *Recovering 3D human pose from monocular images*, in "IEEE Transactions on Pattern Analysis & Machine Intelligence", vol. 28, n° 1, jan 2006, http://lear.inrialpes.fr/pubs/2006/AT06a.

[6] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High dimensional discriminant analysis*, in "Communications in Statistics: Theory and Methods", to appear, http://lear.inrialpes.fr/pubs/2006/BGS06b.

[7] C. BOUVEYRON, S. GIRARD, C. SCHMID. *Class-specific subspace discriminant analysis for high-dimensional data*, Lecture Notes in Computer Science, n° 3940, Springer Verlag, 2006, p. 139–150, http://lear.inrialpes.fr/pubs/2006/BGS06a.

[8] T. BROX, R. VAN DEN BOOMGAARD, F. LAUZE, J. VAN DE WEIJER, J. WEICKERT, P. MRÁZEK, P. KORNPROBST. *Visualization and image processing of tensor fields: adaptive structure tensors and their applications*, chap. I, Springer, 2006, p. 17–47, http://lear.inrialpes.fr/pubs/2006/BVLVWMK06.

[9] P. CARBONETTO, G. DORKÓ, C. SCHMID, H. KUCK, N. DE FREITAS. *A semi-supervised learning approach to object recognition with spatial integration of local features and sementation cues*, in "Towards category-level object recognition", to appear, Springer, http://lear.inrialpes.fr/pubs/2006/CDSKD06a.

[10] M. EVERINGHAM, A. ZISSERMAN, C. K. I. WILLIAMS, L. V. GOOL, M. ALLAN, C. M. BISHOP, O. CHAPELLE, N. DALAL, T. DESELAERS, G. DORKÓ, S. DUFFNER, J. EICHHORN, J. D. R. FARQUHAR, M. FRITZ, C. GARCIA, T. GRIFFITHS, F. JURIE, T. KEYSERS, M. KOSKELA, J. LAAKSONEN, D. LARLUS, B. LEIBE, H. MENG, H. NEY, B. SCHIELE, C. SCHMID, E. SEEMANN, J. SHAWE-TAYLOR, A. STORKEY, S. SZEDMAK, B. TRIGGS, I. ULUSOY, V. VIITANIEMI, J. ZHANG. *The 2005 PASCAL Visual Object Classes Challenge*, in "Selected Proceedings of the first PASCAL Challenges Workshop", LNAI, Springer, 2006, http://lear.inrialpes.fr/pubs/2006/EZKVAMCDDDDEDFGGJKKL.

[11] T. GEVERS, J. VAN DE WEIJER, H. STOKMAN. *Color image processing: methods and applications: color feature detection: an overview*, chap. 9, CRC press, 2006, http://lear.inrialpes.fr/pubs/2006/GVS06.

[12] S. LAZEBNIK, C. SCHMID, J. PONCE. *A discriminative framework for texture and object recognition using local image features*, in "Towards category-level object recognition", to appear, Springer, http://lear.inrialpes.fr/pubs/2006/LSP06a.

[13] J. PONCE, T. L. BERG, M. EVERINGHAM, D. FORSYTH, M. HEBERT, S. LAZEBNIK, M. MARSZAŁEK, C. SCHMID, C. RUSSELL, A. TORRALBA, C. WILLIAMS, J. ZHANG, A. ZISSERMAN. *Dataset issues in object recognition*, in "Towards Category-Level Object Recognition", to appear, Springer, http://lear.inrialpes.fr/pubs/2006/PBEFHLMSRTWZZ06.

[14] F. ROTHGANGER, S. LAZEBNIK, C. SCHMID, J. PONCE. *3D object modeling and recognition from photographs and image sequences*, in "Towards category-Level object recognition", to appear, Springer, http://lear.inrialpes.fr/pubs/2006/RLSP06b.

[15] F. ROTHGANGER, S. LAZEBNIK, C. SCHMID, J. PONCE. *Segmenting, modeling, and matching video clips containing multiple moving objects*, in "IEEE Transactions on Pattern Analysis & Machine Intelligence", to appear, http://lear.inrialpes.fr/pubs/2006/RLSP06a.

[16] F. ROTHGANGER, S. LAZEBNIK, C. SCHMID, J. PONCE. *Object modeling and recognition using local affine-invariant image descriptors and multi-view spatial contraints*, in "International Journal of Computer Vision", vol. 66, n° 3, 2006, http://lear.inrialpes.fr/pubs/2006/RLSP06.

[17] C. SMINCHISESCU, B. TRIGGS. *Fast mixing hyperdynamic sampling*, in "Journal of Image & Vision Computing", Special issue on ECCV'02 papers, vol. 24, n° 3, mar 2006, p. 279–289, http://lear.inrialpes.fr/pubs/2006/ST06.

[18] B. TRIGGS, M. SDIKA. *Boundary conditions for Young - van Vliet recursive filtering*, in "IEEE Transactions on Signal Processing", vol. 54, n° 5, may 2006, http://lear.inrialpes.fr/pubs/2006/TS06.

[19] J. VERBEEK, J. NUNNINK, N. VLASSIS. *Accelerated EM-based clustering of large data sets*, in "Data Mining and Knowledge Discovery", vol. 13, nº 3, nov 2006, p. 291–307, http://lear.inrialpes.fr/pubs/2006/VNV06.

[20] J. VERBEEK. *Learning non-linear image manifolds by combining local linear models*, in "IEEE Transactions on Pattern Analysis & Machine Intelligence", vol. 28, nº 8, aug 2006, p. 1236–1250, http://lear.inrialpes.fr/pubs/2006/Ver06.

[21] J. VERBEEK, N. VLASSIS. *Gaussian fields for semi-supervised regression and correspondence learning*, in "Pattern Recognition", vol. 39, nº 10, oct 2006, p. 1864–1875, http://lear.inrialpes.fr/pubs/2006/VV06.

[22] J. ZHANG, M. MARSZAŁEK, S. LAZEBNIK, C. SCHMID. *Local features and kernels for classification of texture and object categories: a comprehensive study*, in "International Journal of Computer Vision", to appear, http://lear.inrialpes.fr/pubs/2006/ZMLS06a.

[23] J. VAN DE WEIJER, T. GEVERS, A. BAGDANOV. *Boosting color saliency in image feature detection*, in "IEEE Transactions on Pattern Analysis & Machine Intelligence", vol. 28, nº 1, jan 2006, p. 150–156, http://lear.inrialpes.fr/pubs/2006/VGB06.

[24] J. VAN DE WEIJER, T. GEVERS, A. SMEULDERS. *Robust photometric invariant features from the color tensor*, in "IEEE Transactions on Image Processing", vol. 15, nº 1, jan 2006, http://lear.inrialpes.fr/pubs/2006/VGS06.

Publications in Conferences and Workshops

[25] A. AGARWAL, B. TRIGGS. *A local basis representation for estimating human pose from cluttered images*, in "Asian Conference on Computer Vision", jan 2006, http://lear.inrialpes.fr/pubs/2006/AT06.

[26] A. AGARWAL, B. TRIGGS. *Hyperfeatures - multilevel local coding for visual recognition*, in "European Conference on Computer Vision", 2006, http://lear.inrialpes.fr/pubs/2006/AT06b.

[27] C. BOUVEYRON, S. GIRARD, C. SCHMID. *Classification des données de grande dimension : application à la vision par ordinateur*, in "2èmes Rencontres Inter-Associations sur la classification et ses applications", Lyon, France, 2006, http://lear.inrialpes.fr/pubs/2006/BGS06e.

[28] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High dimensional data clustering*, in "17th International Conference on Computational Statistics", 2006, p. 812–820, http://lear.inrialpes.fr/pubs/2006/BGS06c.

[29] C. BOUVEYRON, J. KANNALA, C. SCHMID, S. GIRARD. *Object localization by subspace clustering of local descriptors*, in "5th Indian Conference on Computer Vision, Graphics and Image Processing", 2006, http://lear.inrialpes.fr/pubs/2006/BKSG06.

[30] N. DALAL, B. TRIGGS, C. SCHMID. *Human detection using oriented histograms of flow and appearance*, in "European Conference on Computer Vision", 2006, http://lear.inrialpes.fr/pubs/2006/DTS06.

[31] G. DORKÓ, C. SCHMID. *Maximally stable local description for scale selection*, in "European Conference on Computer Vision", 2006, http://lear.inrialpes.fr/pubs/2006/DS06.

[32] M. HEIKKILA, M. PIETIKAINEN, C. SCHMID. *Description of interest regions with center-symmetric local binary patterns*, in "5th Indian Conference on Computer Vision, Graphics and Image Processing", 2006, http://lear.inrialpes.fr/pubs/2006/HPS06.

[33] D. KNOSSOW, J. VAN DE WEIJER, R. HORAUD, R. RONFARD. *Articulated-body tracking through anisotropic edge detection*, in "Workshop on Dynamical Vision, in conjecture with ECCV", 2006, http://lear.inrialpes.fr/pubs/2006/KVHR06.

[34] D. LARLUS, G. DORKÓ, F. JURIE. *Création de vocabulaires visuels efficaces pour la catégorisation d'images*, in "Reconnaissance des Formes et Intelligence Artificielle", 2006, http://lear.inrialpes.fr/pubs/2006/LDJ06.

[35] D. LARLUS, F. JURIE. *Latent mixture vocabularies for object categorization*, in "British Machine Vision Conference", 2006, http://lear.inrialpes.fr/pubs/2006/LJ06.

[36] S. LAZEBNIK, C. SCHMID, J. PONCE. *Beyond bags of features: spatial pyramid matching for recognizing natural scene categories*, in "IEEE Conference on Computer Vision & Pattern Recognition", 2006, http://lear.inrialpes.fr/pubs/2006/LSP06.

[37] M. MARSZAŁEK, C. SCHMID. *Spatial weighting for bag-of-features*, in "IEEE Conference on Computer Vision & Pattern Recognition", 2006, http://lear.inrialpes.fr/pubs/2006/MS06.

[38] F. MOOSMANN, D. LARLUS, F. JURIE. *Learning saliency maps for object categorization*, in "ECCV International Workshop on The Representation and Use of Prior Knowledge in Vision", 2006, http://lear.inrialpes.fr/pubs/2006/MLJ06.

[39] F. MOOSMANN, B. TRIGGS, F. JURIE. *Randomized clustering forests for building fast and discriminative visual vocabularies*, in "Neural Information Processing Systems (NIPS)", nov 2006, http://lear.inrialpes.fr/pubs/2006/MTJ06.

[40] E. NOWAK, F. JURIE. *Learning visual distance function for object identification from one example*, in "LCE Workshop in conjunction with NIPS'06", 2006, http://lear.inrialpes.fr/pubs/2006/NJ06b.

[41] E. NOWAK, F. JURIE, B. TRIGGS. *Sampling strategies for bag-of-features image classification*, in "European Conference on Computer Vision", Springer, 2006, http://lear.inrialpes.fr/pubs/2006/NJT06.

[42] C. PANTOFARU, G. DORKÓ, C. SCHMID, M. HEBERT. *Combining regions and patches for object class localization*, in "Beyond Patches Workshop in conjunction with CVPR", New York, 2006, http://lear.inrialpes.fr/pubs/2006/PDSH06.

[43] T. QUACK, V. FERRARI, L. V. GOOL. *Video mining with frequent itemset configurations*, in "International Conference on Image and Video Retrieval", 2006, http://lear.inrialpes.fr/pubs/2006/QFV06.

[44] N. SEBE, T. GEVERS, S. DIJKSTRA, J. VAN DE WEIJER. *Evaluation of intensity and color corner detectors for affine invariant salient regions*, in "Beyond Patches Workshop in conjunction with CVPR", 2006, http://lear.inrialpes.fr/pubs/2006/SGDV06.

[45] N. SEBE, T. GEVERS, J. VAN DE WEIJER, S. DIJKSTRA. *Corners detectors for affine invariant salient regions: is color important?*, in "International Conference on Image and Video Retrieval", 2006, http://lear.inrialpes.fr/pubs/2006/SGVD06.

[46] A. THOMAS, V. FERRARI, B. LEIBE, T. TUYTELAARS, B. SCHIELE, L. V. GOOL. *Towards multi-view object class detection*, in "IEEE Conference on Computer Vision & Pattern Recognition", 2006, http://lear.inrialpes.fr/pubs/2006/TFLTSV06.

[47] J. ZHANG, M. MARSZAŁEK, S. LAZEBNIK, C. SCHMID. *Local features and kernels for classification of texture and object categories: a comprehensive study*, in "Beyond Patches Workshop in conjunction with CVPR", 2006, http://lear.inrialpes.fr/pubs/2006/ZMLS06.

[48] Z. ZIVKOVIC, J. VERBEEK. *Transformation invariant component analysis for binary images*, in "IEEE Conference on Computer Vision & Pattern Recognition", 2006, http://lear.inrialpes.fr/pubs/2006/ZV06.

[49] J. VAN DE WEIJER, C. SCHMID. *Blur robust and color constant image description*, in "International Conference on Image Processing", 2006, http://lear.inrialpes.fr/pubs/2006/VS06a.

[50] J. VAN DE WEIJER, C. SCHMID. *Coloring local feature extraction*, in "European Conference on Computer Vision", vol. Part II, Springer, 2006, p. 334–348, http://lear.inrialpes.fr/pubs/2006/VS06.

Internal Reports

[51] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High-dimensional data clustering*, Technical report, n° 1083M, LMC-IMAG, mar 2006, http://lear.inrialpes.fr/pubs/2006/BGS06d.

[52] V. FERRARI, L. FEVRIER, F. JURIE, C. SCHMID. *Groups of adjacent contour segments for object detection*, Technical report, n° 5980, INRIA, sep 2006, http://lear.inrialpes.fr/pubs/2006/FFJS06.

References in notes

[53] D. LOWE. *Distinctive image features from scale-invariant keypoints*, in "International Journal on Computer Vision", vol. 60, n° 2, 2004, p. 91–110.

[54] K. MIKOLAJCZYK, C. SCHMID. *A performance evaluation of local descriptors*, in "IEEE Transactions on Pattern Analysis & Machine Intelligence", vol. 27, n° 10, 2005, p. 1615–1630, http://lear.inrialpes.fr/pubs/2005/MS05.

[55] K. MIKOLAJCZYK, T. TUYTELAARS, C. SCHMID, A. ZISSERMAN, J. MATAS, F. SCHAFFALITZKY, T. KADIR, L. V. GOOL. *A comparison of affine region detectors*, in "International Journal of Computer Vision", vol. 65, n° 1/2, 2005, p. 43–72, http://lear.inrialpes.fr/pubs/2005/MTSZMSKG05.

[56] J. SIVIC, A. ZISSERMAN. *Video Google: A Text Retrieval Approach to Object Matching in Videos*, in "ICCV", vol. 2, oct 2003, p. 1470–1477, http://www.robots.ox.ac.uk/~vgg.