INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team LEAR

Learning and Recognition in Vision

Grenoble - Rhône-Alpes

THEME COG

Table of contents

1. Team

LEAR is a joint team of INRIA and the LJK laboratory, a joint research unit of the Centre National de Recherche Scientifique (CNRS), the Institut National Polytechnique de Grenoble (INPG) and the Université Joseph Fourier (UJF).

Head of the team

Cordelia Schmid [Research Director INRIA (DR2), HdR]

Permanent researchers

Hervé Jégou [Researcher INRIA (CR2)]
Frédéric Jurie [Researcher CNRS (CR1) until 08/2007, now Associated Member, Prof. at Univ. Caen, HdR]
Bill Triggs [Researcher CNRS (CR1), until 10/2007, HdR]
Jakob Verbeek [Researcher INRIA (CR2), from 09/2007, previously post-doc with LEAR]

Faculty members

Roger Mohr [Professor and head at ENSIMAG, HdR]
Laurent Zwald [Associate professor at UJF]

Administrative assistant

Anne Pasteur [Secretary INRIA]

Postdoctoral fellows

Moray Allan [EU project CLASS, 12/2007-12/2008]
Hakan Cevikalp [Techno-Vision project ROBIN, 07/2006-07/2007]
Tingting Jiang [INRIA, 12/2007-08/2009]
Xiaoyang Tan [EU project CLASS, 09/2006-10/2007]
Joost Van de Weijer [Marie Curie fellowship, 04/2005-10/2007]

Technical staff

Matthijs Douze [EU project AceMedia, 01/2005-06/2008]
Yves Gufflet [EU project CLASS, 05/2006-05/2008]
Benoit Mordelet [GRAVIT project, 09/2007-08/2008]
Benjamin Ninassi [Techno-Vision project ROBIN, 02/2005-02/2007]
Christophe Smekens [INRIA, ODL, 09/2007-08/2009]

PhD students

Juliette Blanchet [UJF, MENESR scholarship co-supervised w. INRIA team MISTIS, 10/2004-10/2007]
Christopher Bourez [INPG, EU project CLASS, 01/2006-09/2007]
Christophe Damerval [UJF, MENESR scholarship co-supervised w. MOSAIC team of LMC, from 10/2004]
Matthieu Guillaumin [INPG, ENS Ulm scholarship, from 09/2006]
Hedi Harzallah [INPG, MBDA project, from 02/2007]
Alexander Klaeser [INPG, EU project CLASS, from 11/2006]
Diane Larlus [INPG, MENESR scholarship, from 10/2005]
Joerg Liebelt [INPG, EADS scholarship, co-supervised w. TU Munich, from 01/2007]
Marcin Marszalek [INPG, Marie Curie project VISITOR, from 09/2005]
Eric Nowak [INPG, CIFRE scholarship w. Bertin, from 02/2004]

MSc students

Gagan Gupta [Indian Institute of Technology, 06/2006-06/2007]
Sameh Hamrouni [ENSI Tunis, 02/2007-02/2008]
Sabit Hussain [Supelec Rennes, 02/2007-08/2007]
Josip Krapac [graduated, Univ. Zagreb, 04/2007-09/2007]
Thomas Mensink [graduated, Univ. Amsterdam, 09/2007-02/2008]
Van Hanh Nguyen [IFI Hanoi, 03/2007-06/2007]

Student interns

Julien Barrois [ENSIMAG, 07/2007-09/2007]

Pierre Benard [ENSIMAG, 06/2007-09/2007]
Olivier Schwander [Bachelor ENS Lyon, 06/2007-07/2007]
Francois Visconte [UJF, 09/2007-08/2008]

Visiting scientist

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

2. Overall Objectives

2.1. Introduction

LEAR's main focus is learning based approaches to visual object recognition and scene interpretation, particularly for object category detection, 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 classes, for example based on 2D shape descriptors or 3D object category representations, 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.
- **Recognition in videos.** Recognition in videos adds additional, i.e., temporal, information to visual recognition. The difficulty is to define appropriate temporal descriptors and to integrate them into the recognition framework. Video, furthermore, permits to easily acquire large quantities of image data often associated with text. This data needs to be handled efficiently: we need to develop adequate data structures; text classification can help to select relevant parts of the video. Humans and their activities are one of the most frequent and interesting subjects of videos, but also one of the hardest to analyze owing to the complexity of the human form, clothing and movements. Our research aims at developing robust visual shape descriptors to characterize humans and their movements with little or no manual modeling.

2.2. Highlights of the year

- **Winner of PASCAL VOC 2007 image classification competition**. LEAR participated in the PASCAL Visual Object Classes Challenge 2007, competing against recognition methods developed by the leading academic and industrial teams world-wide. LEAR's approach won the classification contest for 19 of the 20 object classes. See http://www.pascal-network.org/challenges/VOC/voc2007/ for details.
- **Significant number of publications in the major computer vision conferences**. LEAR's scientific results have resulted in 10 scientific publications in the two major computer vision conferences in 2007: 8 papers at CVPR'2007 (acceptance rate: 28%)—3 of them as orals (acceptance rate: 5%)—and 2 papers at ICCV'2007 (acceptance rate: 23%)—1 of them as oral (acceptance rate: 4%). These papers cover a wide range of topics from semantic learning to image retrieval.
- **Platform for fast image search in large databases**. LEAR has developed an image indexing platform that searches in real time for similar images in very large databases. This platform can retrieve corresponding images even if the query image has undergone significant changes, as for example an important scale change. LEAR is currently transferring and testing this platform in an industrial context.

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.

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 parametrized) 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 labeled 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.

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.

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 have developed methods that reliably find humans in still images and videos as well as methods for semi-automatic structuring of personal photo collections. In the EU FP6 project CLASS we investigate methods for visual learning with little or no manual labeling and semantic-level image and video querying. The ANR R2I will investigate how to search conjointly on images and text.

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

Visual (example based) search. The essential requirement here is robust correspondence between observed images and reference ones, despite large differences in viewpoint or malicious attacks of the images. The reference database is typically also large, requiring efficient indexing of visual appearance. Both of these are core technology areas for our team. Visual search is a key component of many applications. One application is navigation through image and video datasets, which is essential due to the growing number of digital capture devices used by industry and individuals. Another application that currently receives significant attention is copyright protection. Indeed, many images and videos covered by copyright are illegally copied on the Internet, in particular on peer-to-peer networks or on the so-called user-generated content sites such as Flickr, YouTube or Dailymotion. The ANR RAFFUT project will investigate the problem of content protection for videos. Another type of application is the detection of specific content from images and videos, which can, for example, be used to package items of an Internet order with a mobile robot. This is the goal of the ANR project RobM@rket.

Automated object detection. Many applications require the reliable detection and localization of one or a few object classes. Examples are pedestrian detection for automatic vehicle control, airplane detection for military applications and car detection for traffic control. Object detection has often to be performed in less common imaging modalities such as infrared and under significant processing constraints. Our industrial project with Bertin is on vehicle classification in infrared images. A potential application is an outdoor defense system where 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. Our industrial project with MBDA is on detecting objects, for example cars, observed from airplanes or missiles. The difficulties are the presence of severe changes of the imaging conditions and the small size of object regions.

5. Software

5.1. Large-scale image indexing: BigImBaz

Participants: Matthijs Douze, Hervé Jégou, Benoit Mordelet, Cordelia Schmid, Christophe Smekens.

LEAR has developed an image search platform which queries for similar images in a very large database: more than one million images. This platform integrates several of LEAR's scientific contributions on image description and large-scale indexing: detectors, descriptors, retrieval using bag-of-words and inverted files, and geometric verification.

The different components integrated in this platform are under copyright protection, i.e., are registered at the Agence pour la Protection des Programmes (APP). They have been licensed to the start-up MilPix, in charge of their commercial exploitation.

5.2. Groups of adjacent contour segments

Participants: Vittorio Ferrari, Frédéric Jurie, Cordelia Schmid.

Pairs of adjacent contour segments (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 portable and includes a README explaining 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 PAS features, please refer to [7] as well as section 6.1.1.

5.3. Histogram of oriented gradient object detection

Participants: Navneet Dalal, Cordelia Schmid, Bill Triggs.

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 [39], 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).

5.4. Extracting and describing interest points

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

Local descriptors [41] computed at affine invariant local regions [42] 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 frequently used in the literature to evaluate new detectors and descriptors.

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 on purpose 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 easily. Additional goals of the library include portability to many platforms and architectures, 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.

According to the sourceforge statistics (see http://sourceforge.net/projects/libit), this library has been downloaded more than 650 times this year.

5.6. Datasets

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

Relevant datasets are important to assess recognition methods. 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, viewpoint changes, changes in illumination, etc. Benchmarking procedures allow to compare the strengths of different approaches. 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, in particular the *Robin dataset* funded by a Techno-Vision grant, see section 8.1.5 for details, but also our own research datasets the *INRIA humans and horses datasets*.

5.6.1. The ROBIN dataset

The Robin dataset includes annotated images as well as performance metrics for multi-class object detection and image categorization. It consists of six different datasets, i.e., several hundred of annotated images. They can be downloaded from the ROBIN website (http://robin.inrialpes.fr). Furthermore, a set of carefully chosen metrics which satisfy the needs expressed by companies, competitors and evaluators is provided, see http://robin.inrialpes.fr/robin_evaluation/downloads/ROBIN_metrics_v6.pdf for additional information. A initial competition took place in July 2007: 10 different teams submitted more than 35 different runs. The results are available at http://robin.inrialpes.fr and have been presented during a workshop organized in Paris in July 2007. An additional competition will be organized in 2008.

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.

5.6.3. IG02 dataset

INRIA has improved the ground truth annotations for the Graz-02 dataset (IG02), a popular natural-scene object category dataset developed by A. Opelt and A. Pinz at Graz University of Technology, see http://www.emt.tugraz.at/~pinz/data/GRAZ_02. This dataset includes images of complex objects (cars, bicycles and people) with high intra-class variability on cluttered backgrounds.

The new annotations created at INRIA (cf. Fig. 1) are object-oriented and more precise. For each object a segmentation mask was drawn, which includes occluded object parts (marked with a different color). Each object was marked "truncated" when it was cut by the image edge, "multiple" when it could not be separated from other objects of this class and "difficult" if it was hard to notice or segment. Images were considered as suitable for training (when there was at least one non-truncated and non-multiple object in the image) and testing (if all the objects in the image could be individually segmented). As those lists had some overlap, we have randomly partitioned it to create a suggested balanced split into a training set and a test set.

6. New Results

6.1. Image descriptors and correspondence

6.1.1. Groups of adjacent contour segments for object detection

Participants: Vittorio Ferrari, Loic Fevrier, Frédéric Jurie, Cordelia Schmid.

To describe shape-based object classes we have introduced in [7] a family of scale-invariant local shape features formed by groups of connected, roughly straight contour segments, and their use for object class detection. 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.

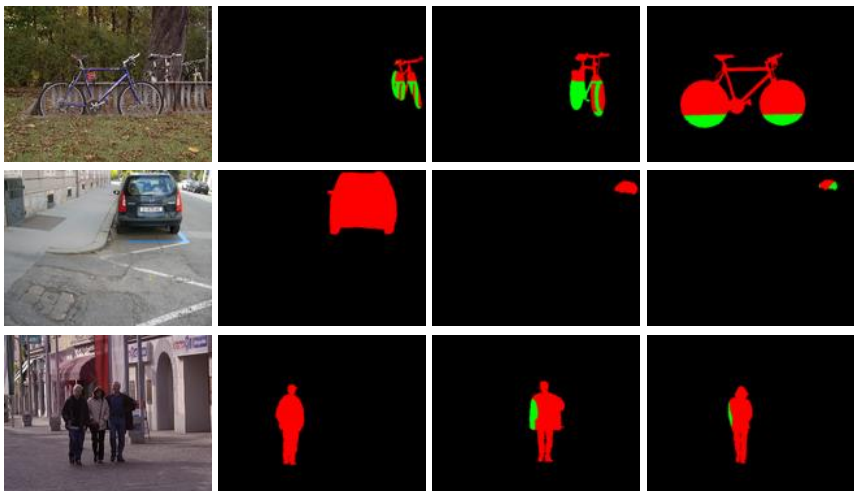


Figure 1. Example images (left column) and segmentation masks for each individual object (right three columns). The resolution of the images is 640x480. Red marks the visible object parts and green is used for occluded object parts.

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

6.1.2. Class-specific features for category-level recognition

Participants: Cordelia Schmid, Tinne Tuytelaars [K. U. Leuven].

The first steps in many class-level object recognition system consist of feature extraction, feature description, and clustering into a visual vocabulary. Our approach [32] 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 quantized, 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 learned 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 work is joint with T. Tuytelaars from K.U. Leuven and was in part realized during her stay with LEAR funded by the INRIA invited professor program.

6.1.3. Image correspondence based on a contextual dissimilarity measure

Participants: Hedi Harzallah, Hervé Jégou, Cordelia Schmid, Jakob Verbeek.

Building on the *Video-Google* image retrieval setup [43], we have designed an enhanced scheme [22] 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 vector. 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.

6.1.4. Top-down color constancy

Participants: Cordelia Schmid, Jakob Verbeek, Joost Van de Weijer.

Color constancy signifies the ability to recognize colors of objects independently of the color of the light source. Most color constancy methods apply a bottom-up approach. Based on an image statistic of low-level image features, these methods compute an estimation of the illuminant color. Our approach [37] investigates the use of high-level visual information for color constancy, see Fig. 2 for an overview of the method. We evaluate a number of illuminant color hypotheses based on the likelihood of their semantic content: is the grass green, the road grey, and the sky blue, in correspondence with our prior knowledge of the world. The semantic likelihood of an image is computed with the probabilistic latent semantic analysis (PLSA) method, which models images as a mixture of semantic topics, such as grass, trees, sky, and building. Based on this semantic likelihood we pick the illuminant which results in the most likely image. We use two approaches to obtain the illuminant hypotheses. The first one uses existing color constancy methods, such as Grey-World, and Max-RGB. The second one casts top-down color constancy hypotheses based on a semantic interpretation of the image and the prior knowledge of the colors of the recognized classes. Experiments show that the use of high-level information improves the illuminant estimation over a purely bottom-up approach.

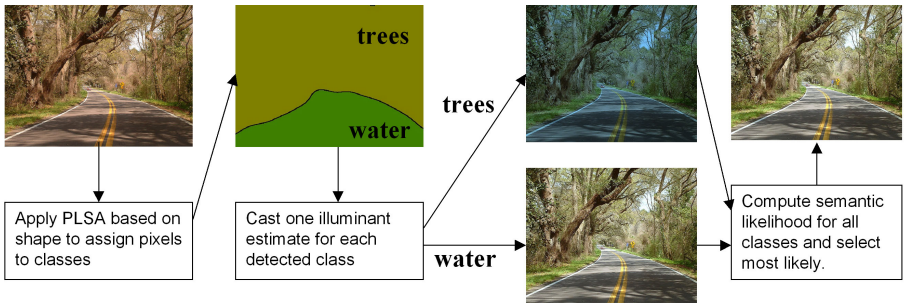


Figure 2. Overview of our color constancy method [37].

6.1.5. Approximate nearest neighbor search

Participants: Laurent Amsaleg [TexMex project-team, INRIA Rennes], Patrick Gros [TexMex project-team, INRIA Rennes], Hervé Jégou, Cordelia Schmid.

It is well known that high-dimensional nearest neighbor retrieval is very expensive and that image description and indexation methods suffer from its computational cost. Significant performance gains can be obtained by using approximate nearest neighbor search methods, as for example Locality-Sensitive Hashing (LSH). In [21] we have proposed an improvement of LSH which performs on-line selection of the most appropriate hash functions from a pool of functions. For this purpose, we have shown that a measure of the expected relevance of a given hash function can be computed without parsing the corresponding hashing cells involved in LSH. An additional improvement originates from the use of E_8 lattices for geometric hashing instead of one dimensional random projections. A performance study based on state-of-the-art high-dimensional descriptors computed on real images shows that our improvements of LSH significantly reduce the search complexity given a level of accuracy. This is joint work with L. Amsaleg and P. Gros from the TexMex project-team at INRIA Rennes.

6.2. Statistical modeling and machine learning for image analysis

6.2.1. Learning distance function for object identification

Participants: Frédéric Jurie, Frank Moosmann, Eric Nowak.

We have proposed a method for comparing instances of never seen objects [12], [28]. The measure is learned from pairs of training images labeled “same” or “different” (see Fig. 3). This is far less informative than the commonly used individual image labels (e.g. “car model X”), but is cheaper to obtain. The proposed algorithm learns the characteristic differences between corresponding local descriptors sampled from pairs of “same” and “different” images (see Fig. 4). These differences are vector quantized by an ensemble of extremely randomized binary trees, and the similarity measure is computed from the quantized differences. The extremely randomized trees are fast to learn, robust due to the redundant information they carry and they have been proved to be very good classifiers. Furthermore, the trees efficiently combine different feature types (SIFT and geometry). The similarity measure is a linear combination of the cluster memberships. Our similarity measure is evaluated on four very different datasets: a dataset with toy cars, a dataset with real cars, a face dataset (a subset of faces in the news) and a dataset with one hundred objects (COIL-100). For all datasets the proposed algorithm largely outperform the state-of-the-art approaches.

Figure 3. Given image 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 be capable of dealing with never seen objects.

6.2.2. Category-level object segmentation

Participants: Frédéric Jurie, Diane Larlus, Eric Nowak.

We have proposed an approach [24] for segmenting objects of a given category, where the category is defined by a set of training images. This problem is also known as the *figure-ground segmentation* problem. 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 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.

More recently we have proposed a fully generative process [25] for object-level segmentation. It combines two complementary approaches. First, a mechanism based on blobs of local regions allows to detect objects using a bag-of-word representation and produces rough image segmentation. Second, a Markov Random Field gives clean cuts and enforces label consistency, guided by local image cues. Gibbs sampling is used to infer

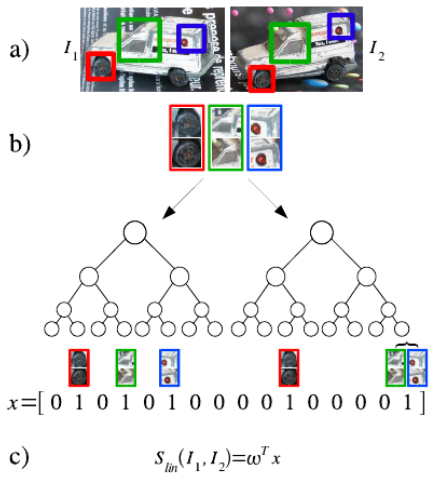


Figure 4. Similarity computation. (a) Detection of corresponding patch pairs. (b) Quantization of these pairs: assignment to clusters using extremely randomized trees. (c) The similarity measure is a linear combination of the cluster memberships.

the model and then labels (objects or background) are produced at the pixel level, giving segmentation masks. This pixel level segmentation can be used to automatically recognize, localize and extract an object from an image (See Fig. 5).

Figure 5. Examples for object detection and segmentation in images. The segmentation masks can be used to extract objects from images.

6.2.3. Face naming from caption-based supervision

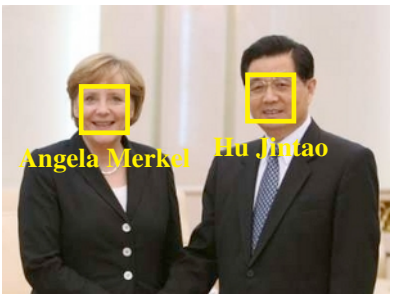
Participants: Yves Gufflet, Matthieu Guillaumin, Thomas Mensink, Cordelia Schmid, Jakob Verbeek.

Over the last decades large digital multimedia archives have appeared, through digitization efforts by broadcasting services, through news oriented media publishing online, or through user provided content concentrated on websites such as YouTube and Flickr. There is a broad ongoing effort to develop methods to disclose

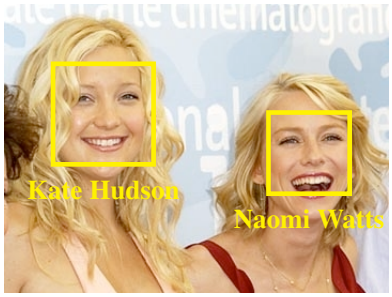
such archives in a user oriented and semantically meaningful way. In particular, there is a great interest in 'unsupervised' systems for automatic content analysis in such archives, that do not require manual annotations to link content to semantic concepts.

We consider two scenarios of naming people in databases of news photos with captions: (i) finding faces of a single person, and (ii) assigning names to all faces. We have combined an initial text-based step, that restricts the name assigned to a face to the set of names appearing in the caption, with a second step that analyzes visual features of faces. By searching for groups of highly similar faces that can be associated with a name, the results of purely text-based search can be greatly improved. Two example images and captions, together with the result of our method, are shown in Fig. 6.

We built upon a recent graph-based approach in which nodes correspond to faces and edges connect highly similar faces. We improved this approach by introducing constraints when optimizing the objective function that limit the method to return at most one face for each name per image, and propose improvements in the low-level methods used to construct the graphs. Furthermore, we have generalized the graph-based approach to face naming in the full data set. In this multi-person naming case the optimization quickly becomes computationally demanding, and we have developed an important speed-up using graph-flows to compute the optimal name assignments in documents. Generative models had previously been proposed to solve the multi-person naming task. We have compared the generative and graph-based methods in both scenarios, and find significantly better performance using the graph-based methods in both cases.



German Chancellor **Angela Merkel** shakes hands with Chinese President **Hu Jintao** (...)



Kate Hudson and **Naomi Watts**, Le Divorce, Venice Film Festival - 8/31/2003.

Figure 6. Typical example images and corresponding captions. In the image we detected faces and associated names are marked, and in the text detected named entities are printed in bold.

6.2.4. Region classification with Markov field aspect models

Participants: Bill Triggs, Jakob Verbeek.

Automatically segmenting an image of a scene into the appearing concepts (such as tree, sky, building, car, road, etc.) is a powerful tool since it not only determines what is present in a scene but also where it appears. Automatic understanding of images in this way can provide useful access to image databases by allowing users to search for images by telling what kind of things should appear where in the image. Existing approaches are mainly based on estimating statistical models on the basis of a collection of manually segmented images. Clearly, manually producing a segmentation is a time consuming task. We have shown [33] 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 labeled images. For the weakly labeled 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 for model estimation 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. In the second step for each concept the associated distributions over image features calculated in the image are re-estimated using the probabilistic segmentation of the first step.

Although aspect models are useful, they ignore the spatial layout 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, a model including spatial dependencies estimated from weakly labeled images yields 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 parameter estimated can be substantially reduced.

6.2.5. Scene segmentation with CRFs learned from partially labeled images

Participants: Bill Triggs, Jakob Verbeek.

As argued in the previous section, automatic segmentation of scenes in the constituent regions and recognizing them is a useful tool for image retrieval. There has been considerable interest in models learned from keywords (as above), or from manually segmented images. However, there is an interesting intermediate option where training images have been partially labeled: some regions in the image have been marked as belonging to a certain concept, e.g. like *building*, and other regions are left unlabeled. This intermediate labeling is easy to use in so called 'generative' probabilistic models, as the one of the previous paragraph. Conditional Random Field (CRF) models learned from completely labeled images are observed to yield state-of-the art segmentation results, but are traditionally learned from completely labeled training images.

In [34] we show how CRFs can be learned from partially labeled training images in a principled manner, and show how resistant the model is to missing training labels. The latter was explored by gradually removing pixels from the training images, as illustrated in Fig. 7, and measuring performance over the model learned from this sequence of increasingly poorer labeling. We found that the model is very robust to missing labeling: when only 40% of the pixels is labeled we obtain essentially the same performance as when 70% of the pixels are labeled. This is possible because when the model is learned, we effectively average over possible ways to complete the partial labeling, in this way label information is propagated to the unlabeled areas. We also explored the effect of so called 'aggregate features' that use the image features of large areas (up to the complete image) to disambiguate local measurements. We find that both the aggregate features and the Markov random field coupling contribute significantly to the recognition accuracy.

6.2.6. Semi-supervised dimensionality reduction using pairwise equivalence constraints

Participants: Hakan Cevikalp, Frédéric Jurie, Alexander Klaeser, Jakob Verbeek.

The standard way in which classification problems are solved is to collect a set of input patterns labeled with the correct class, and then to fit a classification function on these input-output pairs. However, in many applications obtaining labels is a costly procedure as it often requires human effort. On the other hand, in some applications side information –given in the form of pairwise constraints telling that the two input patterns belong to the same class, or to different classes– is available without or with little extra cost. For instance, faces extracted from successive video frames in roughly the same location can be assumed to represent the same person, whereas faces extracted in different locations in the same frame can be assumed to be from different persons.

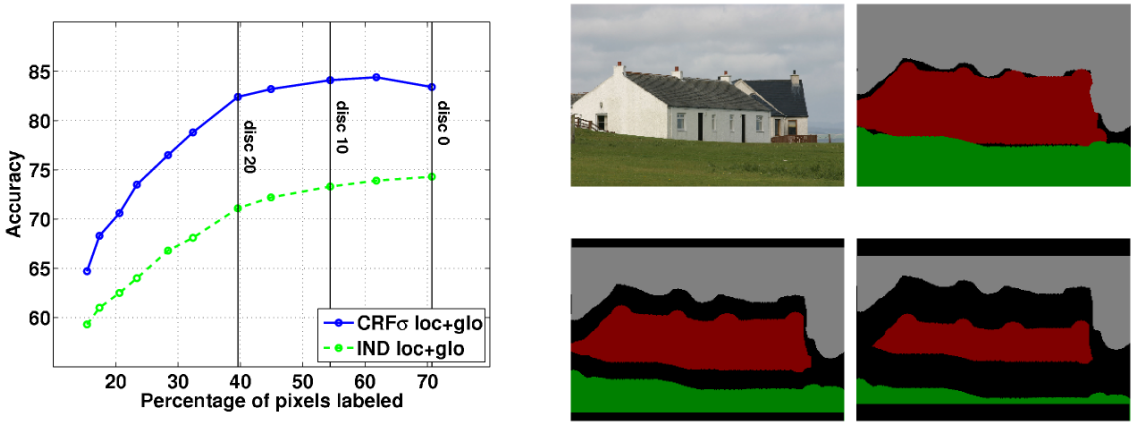


Figure 7. Recognition performance for two of our models when learning from increasingly eroded label images (left). Example image with its original annotation, and erosion thereof with disk of size 10 and 20 (right). Black pixels are not annotated, green pixels correspond to grass, gray pixels to sky, and red pixels to building.

In [18] we propose a semi-supervised dimensionality reduction scheme which uses side information in the form of pairwise equivalence constraints to improve clustering and classification performance. Our algorithm first finds neighboring points for each input to create a weighted neighborhood graph. Then, the side-information constraints are used to modify the neighborhood relations and weight matrix to reflect this weak form of supervision. The optimal projection matrix according to our cost function is then identified by solving for the smallest eigenvalue solutions of a $n \times n$ eigenvector problem, where n is the number of input patterns. Experimental results show that our semi-supervised dimensionality reduction method increases performance of subsequent clustering and classification algorithms, see the illustration in Fig. 8.

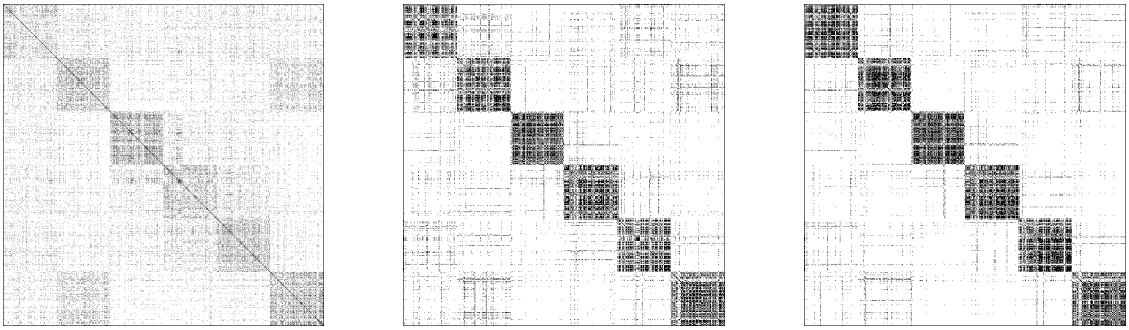
6.3. Visual object recognition

6.3.1. Image classification with bags of local features

Participants: Hedi Harzallah, Svetlana Lazebnik [UNC], Marcin Marszalek, Cordelia Schmid, Joost van de Weijer, Jianguo Zhang, Laurent Zwald.

Many current approaches for image classification are based on visual words (clusters of local descriptors), the per image frequency of these visual words (bag of local features) and a SVM classifier for bags of local features. They obtain excellent results, see Fig. 9 for classification results on the CalTech101 dataset. We have performed a large-scale evaluation of this kind of approach [14]. It showed that the combination of different interest points detectors and descriptors improves the performance. Furthermore, a non-linear χ^2 kernel in combination with a Support Vector Machine (SVM) classifier has shown to outperform several other distances and classifiers. Our bag-of-features approach [14] won the PASCAL VOC challenge on image classification in 2005 and 2006.

This year we have extended the previous approach by adding more image descriptors [38], i.e., by combining different image samplers (dense or sparse), different local descriptors (greylevel or color) and different spatial image grids (global or local). The global image grid defines one bag-of-features per image. Local image grids compute separate bags-of-features per sub-grid and have shown to give excellent results [40]. Given one combination of image sampler, local descriptor and spatial grid, referred to as channel, a histogram of visual words can be computed to represent an image. We can then classify images with a non-linear χ^2 kernel and a SVM. Channels can be combined by either adding the distances of all channels or by selecting weights for



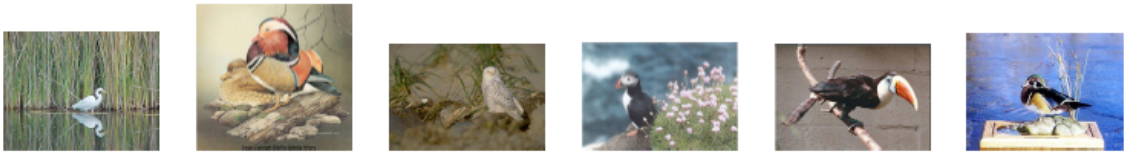


Figure 8. Illustration of the improved representation using our method on a classification task for six bird types. Top: pairwise similarity values (darker is higher) for all image pairs in the original input space, and in representations produced by our methods (using different numbers of pair-wise constraints). Bottom: example images of each of the six classes. Using our method the class structure becomes much clearer in the similarity values.

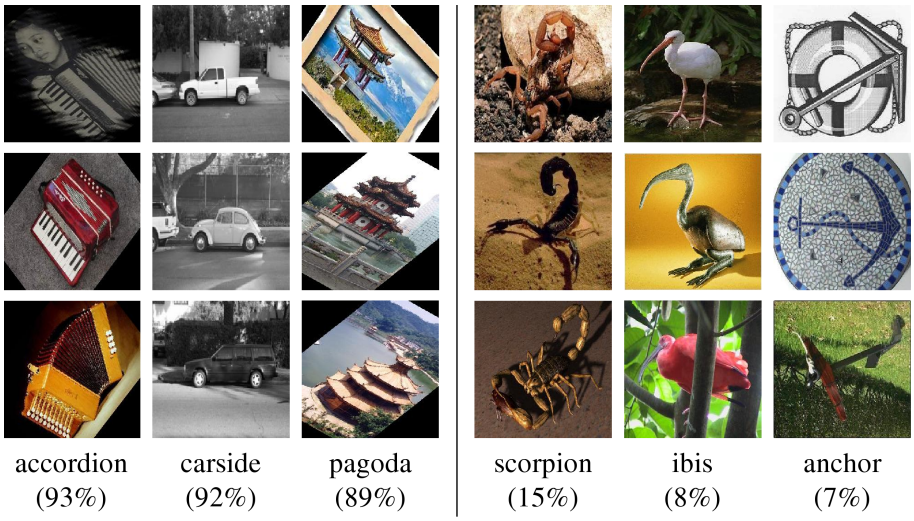


Figure 9. Image examples of the CalTech101 dataset. On the left the three classes with the best classification rates and on the right those with the lowest rates obtained by our approach [14].

each channel with a genetic algorithm. The more sophisticated genetic approach won the classification contest of the PASCAL VOC Challenge 2007, see Fig. 10, a competition with recognition methods developed by the leading academic and industrial teams world-wide.

6.3.2. Accurate object localization with shape masks

Participants: Marcin Marszalek, Cordelia Schmid.

We have developed an approach for object class localization which goes beyond bounding boxes, i.e., it also determines the outline of the objects [26]. Unlike most current localization methods our approach does not require any hypothesis parameter space to be defined. Instead, it directly generates, evaluates and clusters shape masks. Shape masks are aligned based on pairs of matching local features. The similarity of shape masks is measured by the overlap of their surfaces.

Our framework produces a much richer response to object class localization than approaches based on traditional bounding boxes. For example, it easily learns and detects different object viewpoints and articulations, which are often well characterized by the object outline. We evaluate the proposed approach on the challenging natural-scene Graz-02 object classes dataset as well as on the Weizmann horses dataset. The results demonstrate the excellent localization capabilities of our method, cf. Fig. 11 and Fig. 12.

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

Participants: Vittorio Ferrari, Frédéric Jurie, Cordelia Schmid.

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 [19] 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 in cluttered images. As demonstrated by an extensive evaluation, our method can localize object boundaries

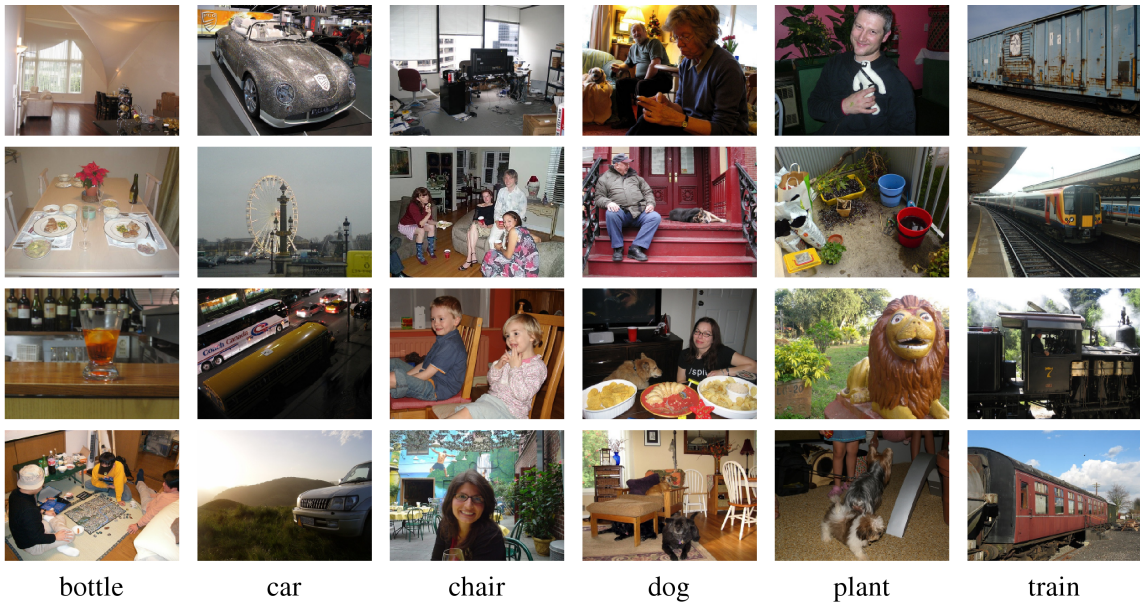


Figure 10. Example images for some classes of the PASCAL *VOC'07 Challenge. Note the difficulty of the task due to intra-class variation, occlusion and background clutter.*

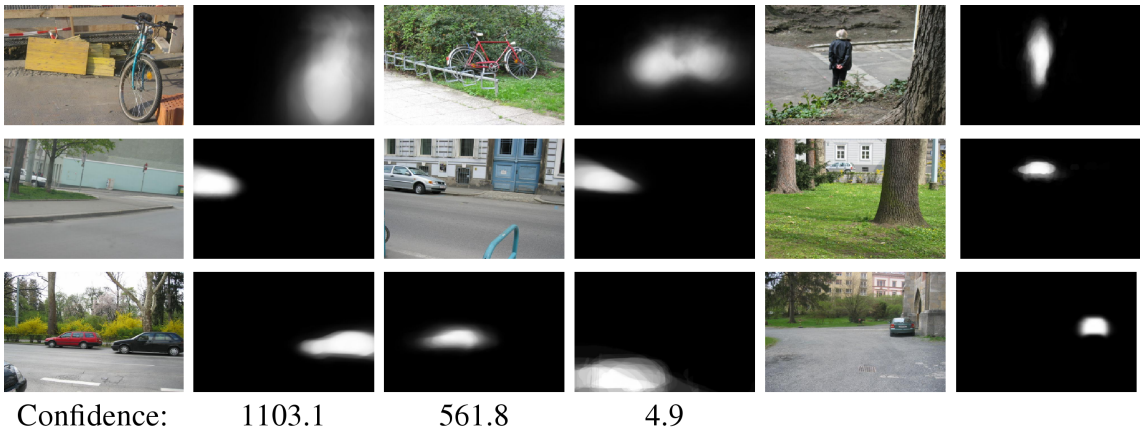


Figure 11. Localization results for the Graz-02 dataset. Note the precise object shape estimations despite occlusion and background clutter. Multiple object instances are detected with subsequent hypotheses as is shown in the bottom row (4 left most columns).

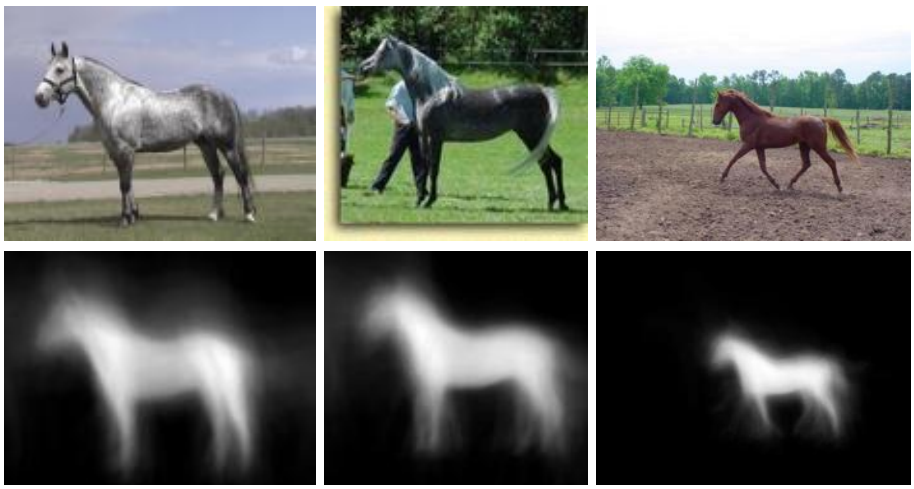


Figure 12. Localization results for the Weizmann horses dataset. Note that the estimated shape is very accurate: the horse articulations are visible.

accurately. Training does not require segmented examples (only bounding-boxes). Figure 13 (top row) shows a few localization results. We can observe the objects are localized very accurately. The bottom row displays two model 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.

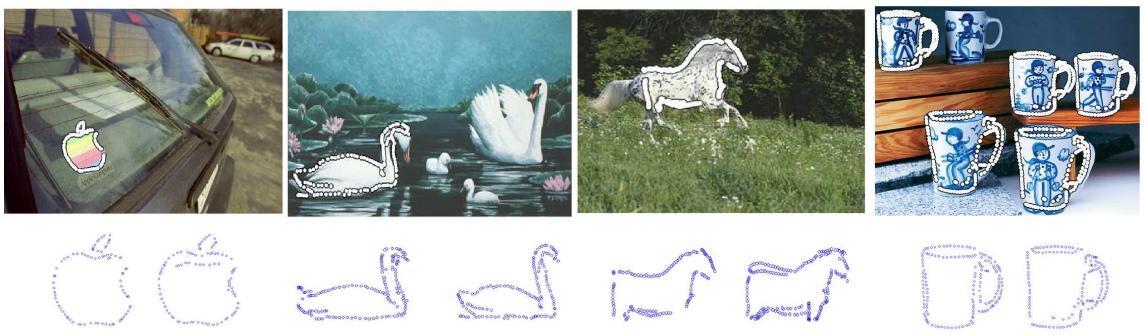


Figure 13. The top row shows a few localization results. Note the accuracy of the detected outlines. The bottom row shows two models for each class—they are learnt from different sets of training images.

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

Participants: Akash Kushal [UIUC], Jean Ponce [ENS Ulm], Cordelia Schmid.

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 [23] for visual object recognition where object classes are represented by graphs of *partial surface*

models (PSMs) obeying loose local geometric constraints. 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 (ENS Ulm); it was partially funded by our associated team Tethys.

6.3.5. *Viewpoint-independent object class detection using 3D feature maps*

Participants: Joerg Liebelt [TU Munich], Cordelia Schmid.

Most existing approaches to viewpoint-independent object class detection combine classifiers for a few discrete views. We propose instead to build 3D representations of object classes which allow to handle viewpoint changes *and* intra-class variability. We do not build a model from 2D features and their geometric constraints. Instead, we resort to a database of existing, fully textured synthetic 3D models, see Fig. 14 (left) for some examples. Our approach renders the synthetic models from different viewpoints and extracts a set of pose and class discriminative features. Discriminative features are obtained by a filtering procedure which identifies features suitable for reliable matching to real image data. A codebook is created based on clusters of these synthetic features encoded by their appearance and 3D position as shown in Fig. 14 (center). During detection local features from real images are matched to the synthetically trained ones as part of a probabilistic voting scheme. Each match casts votes to determine the most likely class and 3D pose of the detected generic object. The most promising votes are then evaluated and refined based on a robust pose estimation step which outputs a 3D bounding box in addition to the 2D localization. Some results are shown in Fig. 14 (right). On a set of calibrated images, we could show that the estimated 3D pose is sufficient to initialize 3D tracking and model registration. On the PASCAL 2006 dataset for motorbikes and cars, the 2D localization results of our approach can compete with state-of-the-art 2D object detectors.

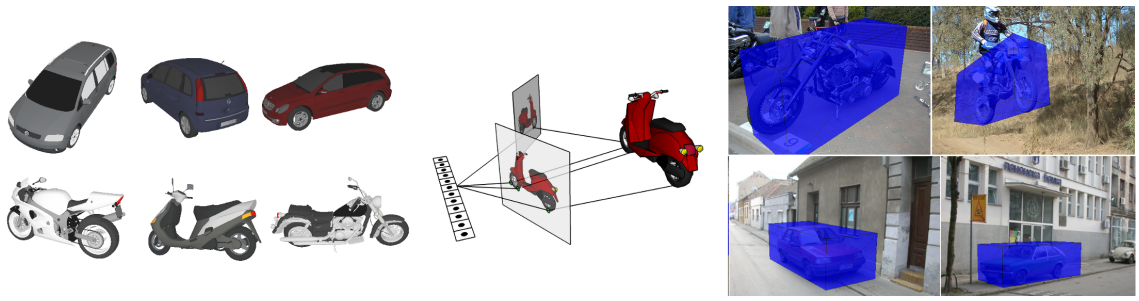


Figure 14. Examples of some 3D models from the training database (left); codebook generation from synthetic views (center); detection results on the PASCAL 2006 database (right).

6.3.6. *Semantic hierarchies for visual object recognition*

Participants: Pierre Benard, Marcin Marszalek, Cordelia Schmid.

In [27] we propose to use a lexical semantic network, i.e., WordNet, to extend the state-of-the-art object recognition techniques. We use the semantics of image labels to integrate prior knowledge about inter-class relationships into the visual appearance learning. We show how to build and train a semantic hierarchy of discriminative classifiers and how to use it to perform object classification. We evaluate how our approach influences the classification accuracy and speed on the PASCAL VOC challenge 2006 dataset, a set of challenging real-world images. We also demonstrate additional features that become available to object recognition due to the extension with semantic inference tools: we can classify high-level categories, such as animals, and we can train part detectors, for example a window detector, by inference in the semantic network.

We also propose to automatically extract vision-oriented semantic information from Flickr. Building on image tags - semantics provided by Flickr users - we show how to construct a rich class hierarchy that reflects visual similarities between classes. In our automatically built class hierarchies we observe semantic relationships similar to the ones present in expert ontologies, but we also discover visual context links and scene-type grouping. Our experiments show the improved performance of our vision-oriented hierarchies over ontology-based hierarchies in terms of modeling the latent visual similarities between object classes.

6.3.7. Learning color names from real-world images

Participants: Cordelia Schmid, Jakob Verbeek, Joost Van de Weijer.

Color names are among the most commonly used visual attributes. Within a computer vision context color naming is the ability to assign linguistic color labels, such as "black", "red" and "mauve", to image pixels. In [36] we propose to learn color names from real-world images retrieved by Google image search, see Fig. 15 for an overview of the method. To learn the color red, we query Google image for "red+color". This process is repeated for a fixed set of color names. The resulting data set is weakly labeled in that we have a single label per image indicating the presence of a color name. Note that since we use all images returned by Google image, the images do not need to contain the labeled color. We apply probabilistic latent semantic analysis (PLSA) to learn the color names. The PLSA model is applied because it allows for multiple "classes" in the same image, which is the case in our Google data set. We model RGB values (words) in images (documents) with mixtures of color names (topics), where mixing weights may differ per image but the topics are shared among all images.

To evaluate the performance of our approach, the color names are tested on two tasks: 1. Retrieval based on color names, where the goal is to rank images according to their resemblance to a color name. This allows users, for example, to search for "red" cars in a car database. 2. Pixel classification, where each pixel is attributed to a color name. The results are compared with color naming based on a 'traditional' approach learned within a laboratory setup. Results show that the color names learned with our method outperform color names learned in a laboratory setup for both retrieval and pixel classification.

We further investigated the usage of color names to describe the color content of local features. Traditionally, the color content is described by photometric invariants. However, they have the drawback that they cannot differentiate between achromatic colors, such as black, grey, and white. A description based on color names can distinguish the achromatic colors, while it also is partially robust to photometric variations. In [35] we compared the two approaches on the task of image classification and found that color names obtained equal or better results than photometric invariants.

6.3.8. Object recognition by integrating multiple image segmentations

Participants: Martial Hebert [CMU], Caroline Pantofaru [CMU], Cordelia Schmid.

The joint tasks of object recognition and segmentation from a single image are complex in their requirement of not only correct classification, but also deciding exactly which pixels belong to the object. Iterating through all possible pixel subsets is prohibitively expensive, leading to recent approaches which use bottom-up unsupervised image segmentation to reduce the size of the configuration space and create better spatial support for features. Bottom-up image segmentation, however, is known to be unstable, with small image perturbations, feature choices, or different segmentation algorithms leading to drastically different image

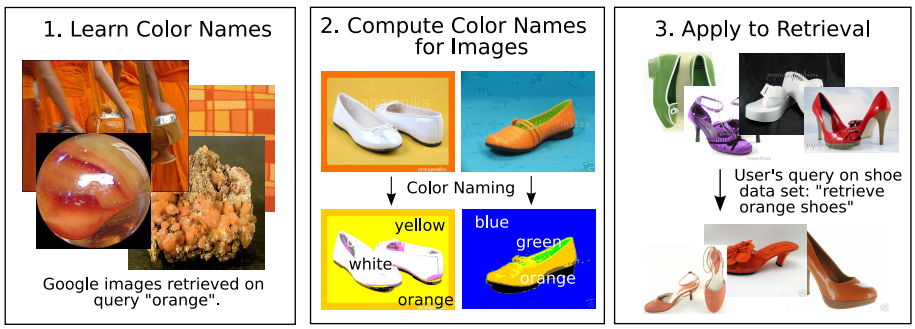


Figure 15. Overview of our method to learn color names from real-world images retrieved by Google [36].

segmentations. This instability has led to advocacy for using multiple segmentations of an image, and makes the method of combining information from these segmentations crucial.

Our approach explores the question of how to best integrate the information from multiple bottom-up segmentations which are created using different scales, algorithms and features. We show how the use of all the bottom-up segmentations in concert leads to improved object segmentation accuracy and creates robustness to outlier image segmentations. Our intuitive formulation shows performance comparable or better than the state of the art on two difficult datasets.

Fig. 16 demonstrates the results of our algorithm. In each row, the images from left to right are: 1) the original image, 2) the ground truth labeled image, 3) the output of our algorithm, 4) an example of a good result from a single image segmentation, and 5) an example of a bad result from a single image segmentation. Notice how the variations between the good and bad single segmentations in columns 4 and 5 are large, but that the combinations of all the segmentations in column 3 performs well. By integrating all of the segmentations our algorithm is robust to the quality of the individual segmentation results. This is joint work with Caroline Pantofaru (CMU) and Martial Hebert (CMU); it was partially funded by our associated team Tethys.

6.4. Recognition in video

6.4.1. Learning realistic human actions from movies

Participants: Ivan Laptev [VISTA project-team, INRIA Rennes], Marcin Marszalek, Cordelia Schmid.

We address recognition of natural human actions in diverse and realistic video settings. This challenging but important subject has mostly been ignored in the past due to several problems, one of which is the lack of realistic and annotated video datasets. Our first contribution is to address this limitation and to investigate the use of movie scripts for automatic annotation of human actions in videos. We evaluate alternative methods for action retrieval from scripts and show benefits of a text-based classifier. Using the retrieved action samples for visual learning, we next turn to the problem of action classification in video. We present a new method for video classification that builds upon and extends several recent ideas including local space-time features, space-time pyramids and multi-channel non-linear SVMs. The method is shown to improve state-of-the-art results on the KTH action dataset by achieving 91.8% accuracy. Given the inherent problem of noisy labels in automatic annotation, we show a high tolerance of our method to annotation errors in the training set. We finally apply the method to learning and classification of challenging action classes in movies and show promising results, see Fig. 17.

6.4.2. Object detection and reconstruction by a humanoid robot

Participants: Frédéric Jurie, Abderrahmane Kheddar [JRL], Diane Larlus, Olivier Stasse [JRL], Kazuito Yokoi [JRL].

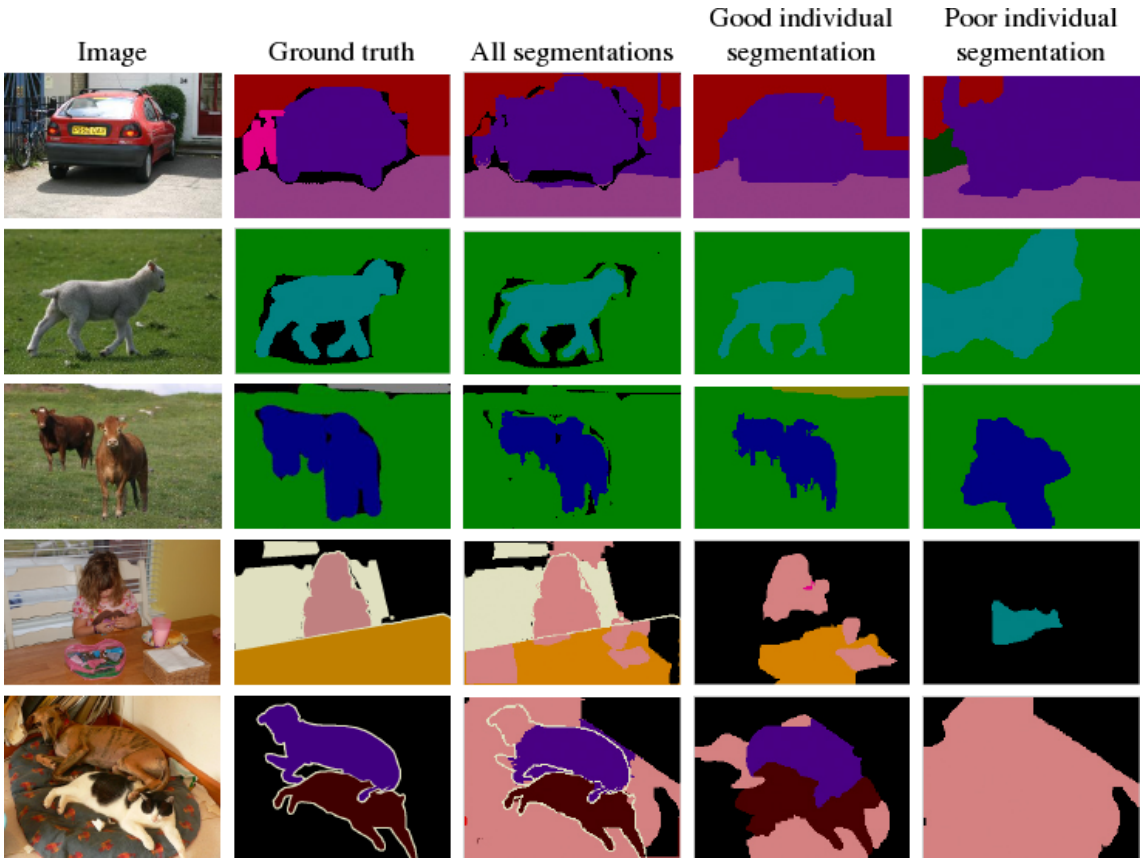


Figure 16. Results of our algorithm which integrates multiple segmentations. The images are, from left to right: 1) the original image, 2) the ground truth labeled image, 3) the output of our algorithm, 4) an example of a good result from a single image segmentation, and 5) an example of a bad result from a single image segmentation.

Figure 17. Example results for action classification trained on automatically annotated data. We show the key frames for test movies with the highest confidence values for true/false positives/negatives.

We have proposed an approach for object detection and reconstruction for visual search by a humanoid robot [29]. This approach is summarized Fig. 18. First a large number of object views are acquired to learn a reliable model of the object. This requires the generation of full body motion including self-collision avoidance for the robot. A next best view algorithm guides the motion and insures that very different views of the object are considered. We, then, introduce a new generative object model which is learned on this set of training images, and applied during testing to achieve object detection. It relies on a patch-based representation which handles efficiently occlusions, scale and pose variations. It also incorporates special features obtained by the stereoscopic vision, such as depth information. A method based on the 3D vision matches some of the local representations previously obtained and allows the robot to precisely estimate the pose of the object. Experiments on the HRP-2 humanoid robotic platform are encouraging.

6.4.3. Human detection and tracking in video sequences

Participants: Alexander Klaeser, Cordelia Schmid.

Our current research includes the detection and tracking of humans in video sequences. We have developed a human detector that is based on body parts (face, head, head+shoulders, see Fig. 19) and exploits knowledge about geometric relations between body parts. To detect body parts we use a soft cascade of higher dimensional features of different types. The features are based on histograms of oriented gradients (HOG), Haar features, and local binary patterns. In order to learn an optimal set of features, we employ a variant of the AdaBoost algorithm. The features are then combined in a cascade which enables fast detection. We have created datasets for training and evaluation of the body part detectors. The model of geometric relations between body parts is learned directly from the training data.

To detect humans in videos we combine a particle filter with our body part detectors. The particle filter increases reliability as well as speed of the detectors. Tracks of human bodies are obtained by first combining body part detections using the human body model. Second, detections in neighboring frames are subsequently merged into tracks. Finally, overlapping tracks are removed. This allows to obtain excellent detection results, see Fig. 19.

6.4.4. Large-scale indexing of videos

Participants: Hervé Jégou, Benoit Mordelet, Cordelia Schmid.

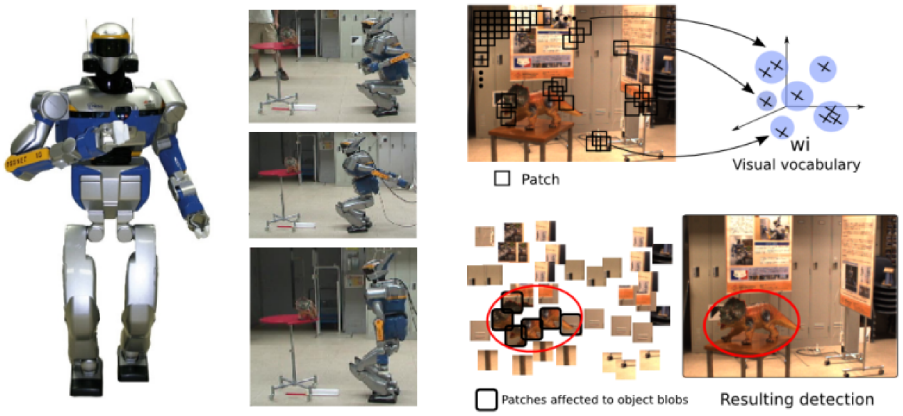


Figure 18. Left: the HRP-2 humanoid robot and three different robot poses obtained by the next best view algorithm. Right: patch based model used for object detection.

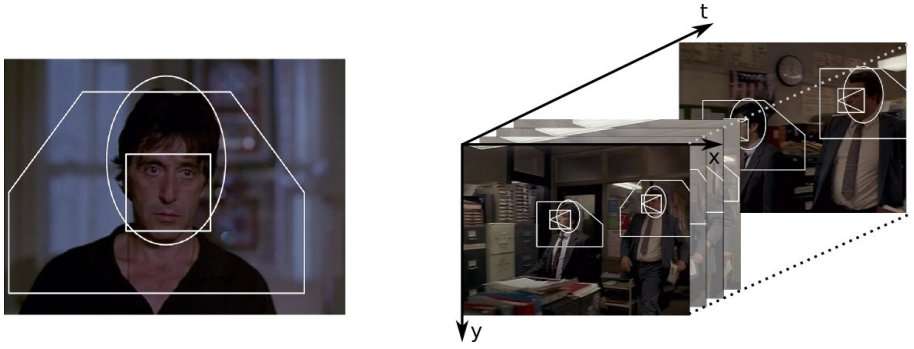


Figure 19. (left) Our human body model uses different body parts: face, head, and head+shoulders; (right) the addition of tracking techniques in videos allows to increase the reliability and speed of our body part detectors.

We have recently addressed the problem of large-scale indexing of videos, i.e., retrieval of videos in the presence of severe image transformations and temporal crops. The indexing system we propose is derived from our image indexing platform BigImBaz. The video description makes use of local descriptors extracted from image regions. A bag-of-word is then used to represent the frames of a given video. At query time geometric verification based on a Hough transform is performed. This amounts to re-ranking the frames according to their geometric consistency.

We are currently extending the video description to take into account temporal video constraints. We evaluate the relevance of spatio-temporal descriptors and of temporal filtering of 2D descriptors. These operations are performed before the bag-of-word computation and do not impact the query time. Evaluation on standard test video databases show the improvement over state-of-the-art techniques.

7. Contracts and Grants with Industry

7.1. Bertin Technologies

Participants: Frédéric Jurie, Roger Mohr, Eric Nowak.

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. Bertin has funded the CIFRE grant for Eric Nowak's PhD thesis, which started in March 2004; the defense will be in March 2008. Bertin Technologies also participates in our Techno-Vision project ROBIN (see paragraph 8.1.5).

7.2. MBDA Aerospatiale

Participants: Hedi Harzallah, Frédéric Jurie, Cordelia Schmid.

We have collaborated with the Aerospatiale section of MBDA for several years. We recall the history of this collaboration: MBDA has funded the PhD of Yves Dufurnaud (1999-2001), a study summarizing the state-of-the-art on recognition (2004) as well as a one year transfer contract on matching and tracking (11/2005-11/2006). In December 2006 we started a three-year contract on object detection in the presence of severe changes of the imaging conditions and if the images of the objects are very small. In particular we will design appropriate image descriptors and use context information to improve the localization performance. The PhD scholarship of Hedi Harzallah which started in February 2007 is funded by this contract. MBDA also participates in our Techno-Vision project ROBIN.

8. Other Grants and Activities

8.1. National Projects

8.1.1. ANR Project GAIA

Participants: Hervé Jégou, Cordelia Schmid.

GAIA is an ANR (Agence Nationale de la Recherche) "blanc" project that is running for 4 years starting from October 2007. It aims at fostering the interaction between three major domains of computer science—computational geometry, machine learning and computer vision—, for example by studying information distortion measures. The partners are the INRIA project-teams Geometrica and LEAR as well as the university of Antilles-Guyane and Ecole Polytechnique.

8.1.2. ANR Project RAFFUT

Participants: Matthijs Douze, Hervé Jégou, Cordelia Schmid.

RAFFUT is an ANR (Agence Nationale de la Recherche) "Audiovisuel et Multimédia" project that started in December 2007 for two years. This project aims at detecting pirated videos. The main issues addressed by this project are 1) how to handle the scalability issues that arise when dealing with extremely large datasets ; 2) how to improve the accuracy of the search if the videos have suffered very strong attacks, as for example low-quality camcorded copies of movies.

The partners are the company Advestigo (http://www.advestigo.com) and LEAR. Advestigo is one of the leaders in the growing "digital asset management market". Its technology is oriented towards video piracy, in particular for detecting fraudulent content on user-generated websites such as YouTube or DailyMotion.

8.1.3. ANR Project R2I

Participants: Moray Allan, Frédéric Jurie, Cordelia Schmid, Jakob Verbeek.

R2I (Recherche d'Image Interactive) is an ANR "Masse de données et connaissances" project that is running for 3 years starting in January 2008. R2I aims at designing methods for interactive image search. These methods will permit to extract semantics from images and transform raw data into a semantically rich representation, to cluster similar images and propose visual summaries of search results, to enable user interaction via semantic concepts related to images and to index very large volumes of images. The final goal of this project is a system for interactive search, which can index about one billion of images and provide users with advanced interaction capabilities. The partners are the company Exalead, a leader in the area of corporate network indexing and a specialist for user-centered approaches, the INRIA project-team Imedia, a research group with a strong background in interactive search of multi-media documents, as well as LEAR and the University of Caen, both specialists in object recognition.

8.1.4. ANR Project RobM@rket

Participants: Frédéric Jurie, Cordelia Schmid.

RobM@rket is an ANR "Systèmes Interactifs et Robotique" project which will start in early 2008 for a two year period. The project aims at developing a robot system which automatically packages the items of an Internet order. The robotic system is a mobile platform with an industrial arm, onto which the different algorithms, i.e., grasping, visual serving and object detection, will be integrated. The partners are the company BA Systèmes, CEA List, the INRIA project-team Lagadic as well as the INRIA project-team LEAR and the University of Caen.

8.1.5. Techno-Vision Project ROBIN

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

LEAR has coordinated the national Techno-Vision project ROBIN, which started in January 2005 for two and a half years. The goal of this project is to measure and compare the performance of visual object recognition systems. The project has developed ground truth datasets and performance metrics. Furthermore, it has run a national competition for comparing different recognition algorithms in July 2007: 10 different teams submitted more than 35 different runs. The results are available at http://robin.inrialpes.fr and have been presented during a workshop organized in Paris in July 2007. An additional competition will be organized in 2008. The project has been funded partly by the French Ministry of Defense and the French Ministry of Research. Several companies and research centers (Bertin Technologies, Cybernetix, DGA, EADS, INRIA, ONERA, MBDA, SAGEM, THALES) as well as 35 public laboratories have participated in the project.

8.1.6. GRAVIT Grant

Participants: Hervé Jégou, Benoit Mordelet, Cordelia Schmid.

The GRAVIT (Grenoble Alpes Valorisation et Innovation Technologique) grant funds transfer and maturation of technology to make it usable in real-world applications. The grant started in May 2007 for a duration of 16 months. Its main goal is to extend our image indexing platform, BigImBaz, to videos. This requires the development of appropriate techniques to deal with large amounts of data inherent in videos. An additional goal is to handle intellectual property issues: (1) ensure that the components used in our platform are not patented and (2) identify at an early stage the components of our systems that should be protected.

8.2. European Projects and Grants

8.2.1. FP6 Integrated Project aceMedia

Participants: Matthijs Douze, Cordelia Schmid, Bill Triggs.

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 project-team 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: Moray Allan, Yves Gufflet, Alexander Klaeser, Cordelia Schmid, Xiaoyang Tan, Bill Triggs, Jakob Verbeek.

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; University of Finland and MPI Tuebingen, Germany.

8.2.3. FP6 Network of Excellence PASCAL

Participants: Juliette Blanchet, Frédéric Jurie, Marcin Marszalek, Cordelia Schmid, Bill Triggs, Jakob Verbeek.

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 Marszalek, 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. VISITOR is funding the PhD of the Polish student Marcin Marszalek, from 09/2005 to 09/2008.

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 started in 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. Associated team Tethys

Participants: David Forsyth [UIUC], Martial Hebert [CMU], Akash Kushal [UIUC], Marcin Marszalek, Caroline Pantofaru [CMU], Jean Ponce [ENS Ulm], Cordelia Schmid.

The associated team Tethys started in January 2007 for one year, and has been recently extended for an additional year (2008). It associates two INRIA project-teams, LEAR and Willow, with two teams in the US, at Carnegie Mellon University and at University of Illinois Urbana-Champaign. The topic of this collaboration is visual recognition of objects with an emphasis on 3D representations for recognition and human activity classification in videos. In 2007, several visits of senior and junior researchers took place, see http://lear.inrialpes.fr/people/schmid/ea_tethys_renvouellement.html for details. A major milestone was a workshop held at UIUC in September which included all of the principals from the U.S. and France teams as well as a number of the students.

8.3.2. JRL (AIST), Tsukuba, Japan

Participants: Frédéric Jurie, Diane Larlus, Olivier Stasse [JRL], Kazuito Yokoi [JRL].

A collaboration between the LEAR project-team and the Japanese-French Robotics Laboratory (JRL), AIST, located in Tsukuba, Japan has started in 2007. D. Larlus has spent two month with a JSPS (Japan Society for the Promotion of Science) funding in Tsukuba. She worked on automatic object model construction and detection in the context of a robotic humanoid platform.

8.3.3. University of Leuven

Participants: Cordelia Schmid, Tinne Tuytelaars [K.U. Leuven].

A collaboration between the LEAR project-team and Tinne Tuytelaars, K.U. Leuven, started in 2006. Tinne visited LEAR twice a month for a few days during 2006 and in the beginning of 2007. These visits were partially funded by the INRIA invited professor program. The collaboration resulted in work on automatic feature selection, published in ICCV'07. We are currently continuing this work, and also working together in the context of the EU project CLASS.

9. Dissemination

9.1. Leadership within the scientific community

- Conference and workshop organization:
  - International Workshop on Object Recognition, 2008 (C. Schmid)
  - Workshop Chair in conjunction with ECCV'2008 (F. Jurie)
- Editorial boards:
  - International Journal of Computer Vision (C. Schmid)
  - IEEE Transactions on Pattern Analysis and Machine Intelligence (B. Triggs)
  - Foundations and Trends in Computer Graphics and Vision (C. Schmid)
- Area chairs:
  - CVPR'07 (C. Schmid)
  - ECCV'08 (C. Schmid)
  - ACCV'07 (C. Schmid)
  - RFIA'08 (C. Schmid)
  - TAIMA'07 steering committee (C. Schmid)
- Program committees:
  - CVPR'07 (H. Jégou, F. Jurie, B. Triggs, J. Verbeek, J. Van de Weijer)
  - ICCV'07 (H. Jégou, F. Jurie, C. Schmid, B. Triggs, J. Verbeek)
  - ACCV'07 (F. Jurie, J. Van de Weijer, J. Verbeek)
  - NIPS'07 (B. Triggs, J. Verbeek)
  - VISAPP'07 (F. Jurie)
  - ACIVS'97 (F. Jurie)
  - RFIA'08 (H. Jégou, F. Jurie)
- Prizes:
  - Winner of the PASCAL VOC 2007 image classification competition. LEAR participated in the PASCAL Visual Object Classes Challenge 2007, competing against recognition methods developed by the leading academic and industrial teams world-wide. LEAR's approach won the classification contest for 19 of the 20 object class categories. See http://www.pascal-network.org/challenges/VOC/voc2007/ for details.
- 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". She participated in several recruitment committees in 2007.
  - C. Schmid is a member of the "conseil de l'Agence d'évaluation de la recherche et de l'enseignement supérieur (AERES)" starting March 2007.
  - C. Schmid has been a member of the evaluation committee for audiovisual and multimedia projects of the Agence Nationale de la Recherche (ANR), 2006 & 2007.

- C. Schmid is a member of the INRIA Grenoble, Rhône-Alpes local scientific committee (bureau du comité des projets).
- C. Schmid has been in charge of international relations at INRIA Grenoble, Rhône-Alpes until September 2007.
- 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 programs of the EU Network of Excellence PASCAL, and co-manages several other programs.

9.2. Teaching

- M. Guillaumin, Practical sessions in computer networking, INPG, 1st year ENSIMAG, 48h
- M. Guillaumin, CAML, INPG, Cycle Préparatoire Polytechnique, 42h
- M. Guillaumin, Tutoring in applied mathematics, INPG, 1st year ENSIMAG, 24H
- H. Jégou, Multi-media databases, INPG, 3rd year ENSIMAG,18h
- F. Jurie, Matching and recognition, INPG, Master IVR, 12h
- F. Jurie, Multi-media databases, INPG, 3rd year ENSIMAG, 18h
- D. Larlus, Practical sessions in functional programming, UPMF, Bachelor 1, 40h
- D. Larlus, Mathematics for computer sciences, UPMF, Master ICA, 24h
- C. Schmid, Matching and recognition, INPG, Master Computer Science, 6h

9.3. Invited presentations

- H. Jégou, *Codage de source robuste et codage conjoint source/canal pour transmission multimedia sur réseaux bruités,* GRETSI, Troyes, September 2007.
- F. Jurie, *Image categorization and image indexing : recent advances*, SETI conference, Mohamedia, Morocco, January 2007.
- F. Jurie, *Randomized clustering forests for image classification*, Pattern Recognition and Computer Vision Colloquium, Prague, Czech Republic, April 2007.
- F. Jurie, *Forêts d'arbres aléatoires pour la construction efficace de vocabulaire visuels discriminants*, Journée ISIS - Signal, Reconnaissance des Formes et Machines à Noyaux, Paris, France, June 2007.
- F. Jurie, *Randomized clustering forests for image classification*, Adaptive Multimedia Retrieval, Paris, France, July 2007.
- D. Larlus, *Automatic object model construction and detection in a robotic framework*, Japanese-French Robotics Laboratory, AIST, Tsukuba, Japan, August 2007.
- M. Marszalek, *Learning object representations for visual object class recognition*, Visual Recognition Challenge Workshop, in conjunction with ICCV'07, Rio de Janeiro, Brazil, October 2007.
- E. Nowak, *Learning visual similarity measures for comparing never seen objects*, Multimodal Interactive Systems Group, TU Darmstadt, Germany, September, 2007.
- E. Nowak, *Learning visual similarity measures for comparing never seen objects*, Computer Graphics and Vision Group, TU Graz, Austria, September, 2007.
- C. Schmid, *Image and object representations for recognition*, 4th International Workshop on Object Categorization, in conjunction with ICCV'07, Rio de Janerio, Brazil, October 2007.
- C. Schmid, *Recognition and matching based on local invariant features*, International Summer School on Computer Vision, Sicily, July 2007.

- C. Schmid, *Beyond bag-of-features: adding spatial and shape information*, 2nd Beyond Patches Workshop, in conjunction with CVPR'07, June 2007.
- C. Schmid, *Beyond bag-of-features: adding spatial and shape information*, Annual Workshop of the Austrian Association for Pattern Recognition (OEAGM'07), Schloss Krumbach, Austria, May 2007.
- C. Schmid, *Beyond bag-of-features: adding spatial and shape information*, Advanced Computer Vision Company, Vienna, Austria, May 2007.
- C. Schmid, *Groups of adjacent contour segments for object detection*, CVPR Area Chair Meeting Workshop, Pittsburgh, U.S., March 2007.
- C. Schmid, *Beyond bag-of-features: adding spatial and shape information*, LIAMA's 10th Anniversary Workshop, Bejing, China, January 2007.

10. Bibliography

Year Publications

Doctoral dissertations and Habilitation theses

[1] J. BLANCHET. *Modèles markoviens et extensions pour la classification de données complexes*, Ph. D. Thesis, Université Joseph-Fourier, Grenoble I, October 2007, http://lear.inrialpes.fr/pubs/2007/Bla07.

Articles in refereed journals and book chapters

[2] A. AGARWAL, B. TRIGGS. *Multilevel image coding with hyperfeatures*, in "International Journal of Computer Vision", to appear, 2008, http://lear.inrialpes.fr/pubs/2007/AT07.

[3] G. BLANCHARD, L. ZWALD. *Finite dimensional projection for classification and statistical learning*, in "IEEE Transactions on Information Theory", to appear, 2008, http://lear.inrialpes.fr/pubs/2008/BZ08.

[4] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High-dimensional data clustering*, in "Computational Statistics and Data Analysis", vol. 52, n^{o} 1, February 2007, p. 502–519, http://lear.inrialpes.fr/pubs/2007/BGS07a.

[5] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High-dimensional discriminant analysis*, in "Communications in Statistics: Theory and Methods", vol. 36, n^{o} 14, January 2007, http://lear.inrialpes.fr/pubs/2007/BGS07.

[6] P. CARBONETTO, G. DORKO, C. SCHMID, H. KUECK, N. DE FREITAS. *Learning to recognize objects with little supervision*, in "International Journal of Computer Vision", to appear, online version available from 07/2007, 2008, http://lear.inrialpes.fr/pubs/2007/CDSKD07.

[7] V. FERRARI, L. FEVRIER, F. JURIE, C. SCHMID. *Groups of adjacent contour segments for object detection*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", to appear, 2008, http://lear.inrialpes.fr/pubs/2008/FFJS08.

[8] H. JÉGOU, C. GUILLEMOT. *Entropy coding with variable length re-writing systems*, in "IEEE Transactions on Communications", vol. 55, n^{o} 3, March 2007, p. 444–452, http://lear.inrialpes.fr/pubs/2007/JG07.

[9] S. MALINOWSKI, H. JÉGOU, C. GUILLEMOT. *Synchronization recovery and state model reduction for soft decoding of variable length codes*, in "IEEE Transactions on Information Theory", vol. 53, n^{o} 1, January 2007, p. 368–377, http://lear.inrialpes.fr/pubs/2007/MJG07.

[10] S. MALINOWSKI, H. JÉGOU, C. GUILLEMOT. *Error recovery properties and soft decoding of quasi-arithmetic codes*, in "EURASIP Journal on Applied Signal Processing", to appear, vol. 2008, 2008, http://www.hindawi.com/GetArticle.aspx?doi=10.1155/2008/752840.

[11] R. MOHR, M. DOUZE, P. STURM. *Géométrie projective, analyse numérique et vision par ordinateur*, in "Bulletin de l'union des professeurs de spéciales", n^{o} 219, July 2007, p. 10–30, http://lear.inrialpes.fr/pubs/2007/MDS07e.

[12] F. MOOSMANN, E. NOWAK, F. JURIE. *Randomized clustering forests for image classification*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", to appear, 2008, http://lear.inrialpes.fr/pubs/2008/MNJ08.

[13] F. ROTHGANGER, S. LAZEBNIK, C. SCHMID, J. PONCE. *Segmenting, modeling, and matching video clips containing multiple moving objects*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", vol. 29, n^{o} 3, mar 2007, p. 477–491, http://lear.inrialpes.fr/pubs/2007/RLSP07.

[14] J. ZHANG, M. MARSZALEK, S. LAZEBNIK, C. SCHMID. *Local features and kernels for classification of texture and object categories: a comprehensive study*, in "International Journal of Computer Vision", vol. 73, n^{o} 2, June 2007, p. 213–238, http://lear.inrialpes.fr/pubs/2007/ZMLS07.

[15] J. V. DE WEIJER, T. GEVERS, A. GIJSENIJ. *Edge-based color constancy*, in "IEEE Transactions on Image Processing", vol. 16, n^{o} 9, September 2007, p. 2207–2214, http://lear.inrialpes.fr/pubs/2007/VGG07.

Publications in Conferences and Workshops

[16] H. CEVIKALP, D. LARLUS, M. DOUZE, F. JURIE. *Local subspace classifiers: linear and nonlinear approaches*, in "IEEE Workshop on Machine Learning for Signal Processing, Thessaloniki, Greece", August 2007, http://lear.inrialpes.fr/pubs/2007/CLDJ07.

[17] H. CEVIKALP, D. LARLUS, F. JURIE. *A supervised clustering algorithm for the initialization of RBF neural network classifiers*, in "IEEE Signal Processing and Communication Applications Conference, Eskisehir, Turkey", June 2007, http://lear.inrialpes.fr/pubs/2007/CLJ07.

[18] H. CEVIKALP, J. VERBEEK, F. JURIE, A. KLAESER. *Semi-supervised dimensionality reduction using pairwise equivalence constraints*, in "International Conference on Computer Vision Theory and Applications", to appear, January 2008, http://lear.inrialpes.fr/pubs/2008/CVJK08.

[19] V. FERRARI, F. JURIE, C. SCHMID. *Accurate object detection with deformable shape models learnt from images*, in "Conference on Computer Vision and Pattern Recognition", June 2007, http://lear.inrialpes.fr/pubs/2007/FJS07.

[20] A. GIJSENIJ, T. GEVERS, J. V. DE WEIJER. *Color constancy by derivative-based gamut mapping*, in "Workshop on Photometric Analysis for Computer Vision in conjuncture with ICCV", October 2007, http://lear.inrialpes.fr/pubs/2007/GGV07d.

[21] H. JÉGOU, L. AMSALEG, C. SCHMID, P. GROS. *Query-adaptative locality sensitive hashing*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing", to appear, April 2008, http://lear.inrialpes.fr/pubs/2008/JASG08.

[22] H. JÉGOU, H. HARZALLAH, C. SCHMID. *A contextual dissimilarity measure for accurate and efficient image search*, in "Conference on Computer Vision and Pattern Recognition", June 2007, http://lear.inrialpes.fr/pubs/2007/JHS07.

[23] A. KUSHAL, C. SCHMID, J. PONCE. *Flexible object models for category-level 3D object recognition*, in "Conference on Computer Vision and Pattern Recognition", June 2007, http://lear.inrialpes.fr/pubs/2007/KSP07.

[24] D. LARLUS, F. JURIE. *Category level object segmentation*, in "International Conference on Computer Vision Theory and Applications", March 2007, http://lear.inrialpes.fr/pubs/2007/LJ07.

[25] D. LARLUS, E. NOWAK, F. JURIE. *Segmentation de catégories d'objets par combinaison d'un modèle d'apparence et d'un champ de Markov*, in "Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle", to appear, January 2008, http://lear.inrialpes.fr/pubs/2008/LNJ08f.

[26] M. MARSZALEK, C. SCHMID. *Accurate object localization with shape masks*, in "Conference on Computer Vision and Pattern Recognition", June 2007, http://lear.inrialpes.fr/pubs/2007/MS07a.

[27] M. MARSZALEK, C. SCHMID. *Semantic hierarchies for visual object recognition*, in "Conference on Computer Vision and Pattern Recognition", June 2007, http://lear.inrialpes.fr/pubs/2007/MS07.

[28] E. NOWAK, F. JURIE. *Learning visual similarity measures for comparing never seen objects*, in "Conference on Computer Vision and Pattern Recognition", June 2007, http://lear.inrialpes.fr/pubs/2007/NJ07.

[29] O. STASSE, D. LARLUS, B. LAGARDE, A. ESCANDE, F. SAIDI, A. KHEDDAR, K. YOKOI, F. JURIE. *Towards autonomous object reconstruction for visual search by the humanoid robot HRP-2*, in "IEEE RAS/RSJ Conference on Humanoids Robot", November 2007, http://lear.inrialpes.fr/pubs/2007/SLLESKYJ07.

[30] X. TAN, B. TRIGGS. *Enhanced local texture feature sets for face recognition under difficult lighting conditions*, in "Analysis and Modelling of Faces and Gestures", LNCS, vol. 4778, Springer, October 2007, p. 168–182, http://lear.inrialpes.fr/pubs/2007/TT07.

[31] X. TAN, B. TRIGGS. *Fusing gabor and LBP feature sets for kernel-based face recognition*, in "Analysis and Modelling of Faces and Gestures", LNCS, vol. 4778, Springer, October 2007, p. 235–249, http://lear.inrialpes.fr/pubs/2007/TT07a.

[32] T. TUYTELAARS, C. SCHMID. *Vector quantizing feature space with a regular lattice*, in "International Conference on Computer Vision", October 2007, http://lear.inrialpes.fr/pubs/2007/TS07.

[33] J. VERBEEK, B. TRIGGS. *Region classification with Markov field aspect models*, in "Conference on Computer Vision and Pattern Recognition", June 2007, http://lear.inrialpes.fr/pubs/2007/VT07.

[34] J. VERBEEK, B. TRIGGS. *Scene segmentation with CRFs learned from partially labeled images*, in "Advances in Neural Information Processing Systems", December 2007.

[35] J. V. DE WEIJER, C. SCHMID. *Applying color names to image description*, in "International Conference on Image Processing", September 2007, http://lear.inrialpes.fr/pubs/2007/VS07.

[36] J. V. DE WEIJER, C. SCHMID, J. VERBEEK. *Learning color names from real-world images*, in "Conference on Computer Vision and Pattern Recognition", June 2007, http://lear.inrialpes.fr/pubs/2007/VSV07.

[37] J. V. DE WEIJER, C. SCHMID, J. VERBEEK. *Using high-level visual information for color constancy*, in "International Conference on Computer Vision", October 2007, http://lear.inrialpes.fr/pubs/2007/VSV07b.

Miscellaneous

[38] M. MARSZALEK, C. SCHMID, H. HARZALLAH, J. V. DE WEIJER. *Learning object representations for visual object class recognition*, Visual Recognition Challenge workshop, in conjunction with ICCV, October 2007, http://lear.inrialpes.fr/pubs/2007/MSHV07.

References in notes

[39] 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.

[40] S. LAZEBNIK, C. SCHMID, J. PONCE. *Beyond bags of features: spatial pyramid matching for recognizing natural scene categories*, in "Conference on Computer Vision and Pattern Recognition", vol. II, June 2006, p. 2169–2178, http://lear.inrialpes.fr/pubs/2006/LSP06.

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

[42] 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^{o} 1/2, 2005, p. 43–72, http://lear.inrialpes.fr/pubs/2005/MTSZMSKG05.

[43] J. SIVIC, A. ZISSERMAN. *Video Google: A Text Retrieval Approach to Object Matching in Videos*, in "IEEE International Conference on Computer Vision", vol. 2, oct 2003, p. 1470–1477, http://www.robots.ox.ac.uk/~vgg/publications/html/sivic06c-abstract.html.