Project/Team LEAR : Learning and Recognition in Vision

Frédéric Jurie, Cordelia Schmid, Bill Triggs

Project-Team: LEAR

Learning and Recognition in Vision
Theme COG
Rhone-Alpes

January 10, 2006

Contents

LEAR is part of the GRAVIR-IMAG laboratory, a Joint Research Unit of INRIA, the Centre National de Recherche Scientifique (CNRS), the Institut National Polytechnique de Grenoble (INPG) and the Université Joseph Fourier (UJF).

1 Team

Head of project team

Cordelia Schmid [DR2, INRIA]

Deputy-head and scientific co-director

Bill Triggs [CR1, CNRS]

Permanent researchers

Frédéric Jurie [CR1, CNRS]

Faculty members

Roger Mohr [Professor, ENSIMAG]

Administrative assistant

Anne Pasteur

Postdoctoral fellows

Vittorio Ferrari [EADS postdoc, 10/2005-10/2006]

Joost Van de Weijer [Marie Curie postdoc, 11/2005-11/2007]

Jakob Verbeek [INRIA postdoc, 12/2005-12/2006]

Jianguo Zhang [ACI MoviStar postdoc, 12/2003-07/2005]

Technical staff

Julien Bohne [MBDA grant, 10/2005-10/2006]

Matthijs Douze [EU projects LAVA & aceMedia, 01/2005-01/2007]

Benjamin Ninassi [Techno-vision project ROBIN, 02/2005-02/2006]

Michaël Sdika [EU project LAVA & aceMedia, 09/2003-02/2005]

PhD students

Ankur Agarwal [INPG from 10/2004, MENESR scholarship]

Juliette Blanchet [UJF from 10/2004, MENESR scholarship co-supervised with INRIA project MISTIS]

Guillaume Bouchard [UJF defended 05/2005, EU project LAVA until 11/2004]

Charles Bouveyron [UJF from 10/2003, MENESR scholarship co-supervised with INRIA project MISTIS]

Christophe Damerval [UJF from 10/2004, MENESR scholarship co-supervised with MOSAIC team of LMC]

Navneet Dalal [INPG from 10/2003, EU project aceMedia]

Gyuri Dorkó [INPG from 10/2003, EU project LAVA]

Diane Larlus [INPG from 10/2005, MENESR scholarship]

Marcin Marszalek [INPG from 9/2005, Marie Curie project VISITOR]

Eric Nowak [INPG from 02/2004, CIFRE scholarship from Bertin]

Juho Kannala [Oulu University, Finland, visiting student 09/2005-12/2005]

Caroline Pantofaru [Carnegie Mellon University, US, visiting student 09/2005-12/2005 on Marie Curie project VISITOR]

MSc students

Diane Larlus [Master INPG IVR, 04/2005-09/2005]

Salil Jain [Master INPG IVR, Embassy scholarship 06/2003-08/2004 then INRIA scholarship until 08/2005]

Marcin Marszalek [Master Warsaw University of Technology, 04/2005-08/2005]

Frank Moosmann [Master Karlsruhe University, 10/2005-04/2006]

Student interns

Pranay Jain [IIT Delhi, India, INRIA internship 05/2005-07/2005]

Subhransu Maji [IIT Kanpur, India, INRIA internship 05/2005-07/2005]

2 Overall Objectives

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

LEAR's main research areas are:

- **Image features and descriptors and robust correspondence.** Many efficient lighting and viewpoint invariant image descriptors are now available, such as affine-invariant interest points and histogram of oriented gradient appearance descriptors. Our current research aims at extending these techniques to give better characterizations of visual object and texture classes and 2D and 3D shape information, and at defining more powerful measures for visual salience, similarity, correspondence and spatial relations.

- **Statistical modelling and machine learning for visual recognition.** Our work on statistical modelling 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. Detectors, datasets as well as evaluation procedures are available at http://lear.inrialpes.fr/software.

- **Visual recognition and content analysis.** Visual recognition requires the construction of exploitable visual models of particular objects and of object and scene categories. Achieving good invariance to viewpoint, lighting, occlusion and background is challenging even for exactly known rigid objects, and these difficulties are compounded when reliable generalization across object categories is needed. Our research combines advanced image descriptors with learning to provide good invariance and generalization. Currently the selection and coupling of image descriptors and learning techniques is largely done by hand, and one significant challenge is the automation of this process, for example using automatic feature selection and statistically-based validation diagnostics.

- **Human detection and activity analysis.** Humans and their activities are one of the most frequent and interesting subjects of images and videos, but also one of the hardest to analyze owing to the complexity of the human form, clothing and movements. Our research in this area uses machine learning techniques and robust visual shape descriptors to characterize humans and their movements with little or no manual modelling. Particular focuses include robust human detection in images and videos, and reconstructing 3D human movement from monocular images.

3 Scientific Foundations

3.1 Image features and descriptors and robust correspondence

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

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

- **Locality and redundancy:** For resistance to variable intra-class geometry, occlusions, changes of viewpoint and background, and individual feature extraction failures, descriptors should have relatively small spatial support and there should be many of them in each image. Schemes based on collections of image patches or fragments are more robust and better adapted to object-level queries than global whole-image descriptors. A typical scheme thus selects an appropriate set of image fragments, calculates robust appearance descriptors over each of these, and uses the resulting collection of descriptors as a characterization of the image or object (a "bag of features" approach – see below).

- **Photometric and geometric invariance:** Features and descriptors must be sufficiently invariant to changes of illumination and image quantization and to variations of local image geometry induced by changes of viewpoint, viewing distance, image sampling and by local intra-class variability. In practice, for local features geometric invariance is usually approximated by invariance to Euclidean, similarity or affine transforms of the local image.

- **Repeatability and salience:** Fragments are not very useful unless they can be extracted reliably and found again in other images. Rather than using dense sets of fragments, we often focus on local descriptors based at particularly salient points – "keypoints" or "points of interest". This gives a sparser and thus potentially more efficient representation, and one that can be constructed automatically in a preprocessing step. To be useful, such points must be accurately relocalizable in other images, with respect to both position and scale.

- **Informativeness:** Notwithstanding the above forms of robustness, descriptors must also be informative in the sense that they are rich sources of information about image content that can easily be exploited in scene characterization and object recognition tasks. Images contain a lot of variety so high dimensional descriptions are required. The useful information should also be manifest, not hidden in fine details or obscure high-order correlations. In particular, image formation is essentially a spatial process, so relative position information needs to be made explicit, e.g. using local feature or context style descriptors rather than global moments or Fourier descriptors.

Partly owing to our own investigations [12, 13, 16], 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 feature" 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. Many of the methods discussed below are related to it.

Our current research on local features is focused on creating detectors and descriptors that are better adapted to particular sensors or particular kinds of imagery, 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.

Another class of image features that we have recently developed for visual object and part detection are dense grids of well-normalized histograms of oriented image gradients. Although less local and less invariant than keypoint based features, such grids are both highly informative and very resistant to photometric variations, making them an excellent choice for robust object detection based on learned 'exemplars' or generalized templates. Our methods based on this approach currently appear to be the best detectors available for standing or walking humans [33].

3.2 Statistical modelling 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.

- 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 pLSA and LDA, and manifold methods such as Isomap/LLA.
- To capture the shapes of complex probability distributions over high dimensional descriptor spaces, we either fit mixture models and similar structured semi-parametric probability models, or reduce them to histograms using vector quantization techniques such as K-means or latent semantic structure models.
- Missing data is common owing to unknown class labels, feature detection failures, occlusions and intra-class variability, so we need to use data completion techniques based on variational methods, belief propagation or MCMC sampling.
- Weakly labelled data is also common – for example one may be told that a training image contains an object of some class, but not where the object is in the image – and variants of

unsupervised, semi-supervised and co-learning are useful for handling this. In general, it is expensive and tedious to label large numbers of training images so less supervised data mining style methods are an area that needs to be developed.

- On the discriminative side, machine learning techniques such as Support Vector Machines, Relevance Vector Machines, and Boosting, are used to produce flexible classifiers and regression methods based on visual descriptors.
- Visual categories have a rich nested structure, so techniques that handle large numbers of classes and nested classes are especially interesting to us.
- Images and videos contain huge amounts of data, so we need to use algorithms suited to large-scale learning problems.

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

3.3 Visual recognition and content analysis

Current progress in visual recognition shows that combining advanced image descriptors with modern learning and statistical modelling 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 (see §3.1). Trying to learn everything from scratch using simpler, non-invariant features would require far too much data: good learning can not 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.

3.4 Human detection and activity analysis

One special case of the above is detecting and recognizing humans, tracking and reconstructing their motions, and recognizing their activities. The importance of humans as subject matter and the complexities of their forms, appearances and motions warrant a special effort in this area. Our current research focuses on two sub-domains. The first is achieving reliable detection and body-part labelling

of humans in images and videos despite changes of viewpoint (from long shot to close up), lighting, clothing and pose. The second is "markerless monocular motion capture" – using learning based models to provide reliable, accurate reconstruction of 3D human pose and motion from unmarked humans in monocular images and videos. Future research will extend this to the classification and analysis of human actions.

4 Application Domains

A solution to the general problem of visual recognition and scene understanding would 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). Our recent past projects related to this area include the EU FP5 projects VIBES (visual correspondence and indexing that quickly finds all occurrences of any given specific object or scene element in a feature-film length video, human motion reconstruction) and LAVA (learning based methods and visual recognition applications suitable for use with mobile devices such as telephones with cameras). In the EU FP6 project AceMedia we are currently developing methods that reliably find humans in still images and videos and categorize their actions and we will shortly begin work on semi-automatic structuring of personal photo collections. A new FP6 project CLASS will investigate methods for visual learning with little or no manual labelling and semantic-level image and video querying.

Human computer interfaces. Our human detection and human motion understanding methods were initially developed for image and video understanding applications, but they are potentially also useful for HCI applications and intelligent environments. This is an area that needs to be developed in the future. A related area in which we have already attempted some technology transfer is **markerless human motion capture** for film and video production.

Visual (example based) navigation. The essential requirement here is robust correspondence between observed images and reference (map) ones, despite large differences in viewpoint. The refer-

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

ence database is typically also large, requiring efficient indexing of visual appearance. Both of these are core technology areas for our team. There are applications to pedestrian and driver aids and to autonomous vehicles including civilian (e.g. hospital robot) and aerospace and military ones.

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

5 Software

5.1 Robust image correspondence and rapid recovery of specific objects and scene elements in image databases

Participants: Matthijs Douze, Michael Sdika, Salil Jain, Gyuri Dorko, Cordelia Schmid, Bill Triggs, Roger Mohr.

Local descriptors [12] based on affine invariant local regions [13] 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. The basic method extracts invariant local descriptors in each image and stores them in a high-dimensional spatial data structure to allow efficient indexing and recovery of similar descriptors. Our current methods use modified KD-tree based data structures, combining KD-tree based voting schemes with a verification step to find similar images in the database.

This technology has been used in two demonstrators developed over the last two years. One is designed for interactive use. It finds the images containing any given object or scene element in a database containing 500 images in about a third of a second. The search object is defined by giving a sample image, for example from a webcam. Images can also be added to the database on the fly, allowing the system to be used for, e.g. vision based navigation. Fig. 1 shows an example of the interface. The method was demonstrated at the 2005 “Fête de la Science” and at the 2005 International Conference on Computer Vision in Beijing.

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

5.2 Visual Localization Demonstrator

Participants: Matthijs Douze, Bill Triggs, Cordelia Schmid, Peter Sturm [MOVI team].

Another use of our viewpoint-invariant image matching technology is to find correspondences between collections of images of the same scene. This allows the camera to be located with respect to the scene, and it is a first step towards building a mosaic or a reconstruction of the scene. Mainly as a test of our matching technology, we participated in the Computer Vision Contest at the 2005 International Conference on Computer Vision. The aim was to make a program that inputs photographs

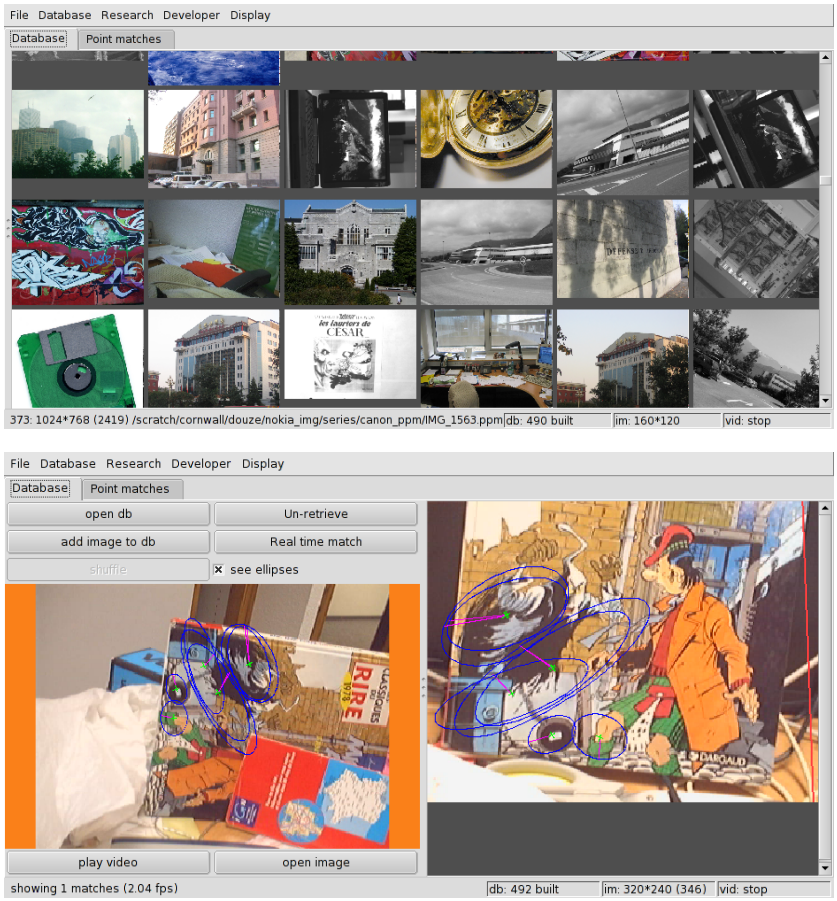


Figure 1: Top row: some examples from an image database, including outdoor scenes, drawings and objects. Bottom row: the matched affine-invariant regions and the estimated transformation between a query and the retrieved image. Note that the same regions (regions of the same appearance) are detected in each image despite the occlusions and viewpoint changes. This allows rich viewpoint-invariant descriptors to be calculated.

of one or more scenes, some of which are labeled with known camera locations (measured by a GPS receiver), and that outputs estimates of the locations of all of the other cameras.

Our entry works as follows. Following robust correspondence, images that were taken from the same viewpoint are found by estimating homographies using RANSAC voting. These image groups are mosaiced, the mosaics are matched with one another and a RANSAC based 5 point resection method is used to compute the relative locations of the cameras for pairs of mosaics with intersecting fields of view. By propagating this information across sets of images, the locations and orientations of the unknown cameras can be recovered based on the known ones. Our method was by far the fastest in the competition owing to our rapid correspondence indexing. It received an honorable mention but further work is required as it only ranked 5th among the 15 competitors on precision, mainly owing to the lack of a final bundle adjustment phase.

5.3 Histogram of Oriented Gradient Object Detection Toolkit

Participants: Navneet Dalal, Bill Triggs, Cordelia Schmid.

As part of the European Union FP6 Integrated Project aceMedia we have developed a toolkit for detecting specific visual object classes such as humans, cars and motorbikes in static images. Although developed originally for human detection [33], 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://pascal.inrialpes.fr/soft/olt) since September 2005. Using this toolkit, we also won 6 of the 10 visual object localization challenges proposed during the European Union Network of Excellence PASCAL's Visual Recognition Challenge.

The software is under copyright protection, registered at the Agence pour la Protection des Programmes (APP) 249, rue de Crie-75019 Paris, France [48].

6 New Results

6.1 Image description and correspondence

Participants: Gyuri Dorko, Matthijs Douze, Vittorio Ferrari, Caroline Pantofaru, Michael Sdika, Joost Van de Weijer, Salil Jain, Christophe Damerval, Cordelia Schmid, Frédéric Jurie, Bill Triggs, Krystian Mikolajczyk [Surrey], Tijmen Moerland [Leiden].

Keywords: feature detection, photometric invariants, grey-level descriptors, shape features, performance evaluation.

6.1.1 Performance evaluation of local detectors and descriptors

We have developed scale- and affine-invariant salient point detectors that give excellent performance for recognizing both specific objects and scenes, and texture and object classes [16]. A performance evaluation has shown that the points and their regions can be detected repeatably in the presence of significant scale changes (up to a factor 4) and affine deformations (viewing angle changes of up to 70 degrees). Various other approaches for detecting affine-invariant interest points or regions have been developed at Leuven, Oxford and Prague universities. We have collaborated with them on a comparison of these approaches. The results [13] show that the different detectors all perform well in the presence of large viewpoint changes. The detectors are complementary and ideally several of them should be used in parallel. None of them outperforms all of the others over all types of scenes and transformations. In most of our experiments, either the Prague detector (called MSER) or our "Hessian-Affine" detector provide the best repeatability score. Another contribution of our study is the carefully designed test setup.

Given a set of stably detected local image regions, we can calculate local image descriptors based on them and use these for matching and recognition. The descriptors should be distinctive and at the same time robust, both to changes in illumination and viewing conditions and to inaccuracies of the region detector. Many different descriptors have been proposed in the literature, and it was unclear

which were the most appropriate for particular problems and how their performance depended on the detector. To help to clarify this, we evaluated the pairing of a number of different interest point detectors with a number of different image descriptors [12]. The evaluation was carried out for various types of images and transformations, using recall/precision as the main quality criterion. By varying the value of the similarity threshold for declaring a match between two descriptors, we generated the curves of the trade-off between the number of correct matches and the number of false matches obtained for an image pair. The ranking of the evaluated descriptors was mostly independent of the interest point detector used, with the SIFT based descriptors [Low04b] performing best. Their success can be explained by their robustness to localization errors and small geometric distortions.

6.1.2 Maximally Stable Local Description for Scale Selection

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

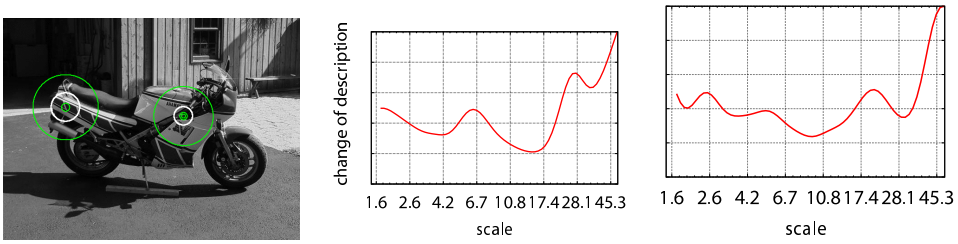


Figure 2: Two examples of our descriptor based scale selection method. The graphs show the change of the local description as a function of scale, respectively for the left and right points. The scales at which the functions have local minima are shown in the image. The bright thick circles corresponds to the global minimum for the point.

Our feature detector has two stages. An interest point detector is run at multiple scales to determine informative and repeatable locations, then for each position we apply a descriptor-based scale selection algorithm to identify maximally stable representations. The chosen scales are the ones at which the descriptor (here SIFT [Low04b]) *changes most slowly* as a function of scale. Fig. 2 illustrates the method on two chosen Harris points. The two functions show how the descriptors change as we increase the scale (the radius) around the two keypoints. The minima of the functions determine the scales at which the descriptions are most stable. Their corresponding regions are depicted by circles in the image. The algorithm selects the *absolute minimum* for each point (shown as a bright circle), but in cases where matching across extreme scale changes is needed we output all minima to provide multiple possible scale selections.

[Low04b] D. G. LOWE, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision 60*, 2, 2004, p. 91–110.

We call this algorithm *Maximally Stable Local Description*. It performs well under changes of viewpoint and lighting conditions, often giving better repeatability than other state-of-the-art methods. In our tests on object category classification it achieved similar or better results on four different datasets while for texture classification it always outperformed existing detectors. Stable orientation estimation (based on the dominant gradient at the keypoint location) can also be integrated into the method and again improves the texture results. Detailed results can be found in [34].

6.1.3 Enriching Local Descriptors with Color Information

Local invariant descriptors are an efficient tool for scene representation due to their robustness with respect to occlusion and geometrical transformations. They are typically computed in two steps, detection of salient and sufficiently invariant local features, followed by the extraction of a robust visual appearance descriptor at the feature location and scale – see fig. 3. To be maximally informative, the descriptor should capture both the visual shape and the color characteristics of the local image region. A considerable amount of research has been dedicated to robust local shape descriptors, the SIFT descriptor being the current reference. The robust description of color has achieved comparatively little attention.

Color is an indispensable characteristic of the visual world, but like grey-levels, absolute color values are not meaningful in themselves as illuminant color, viewpoint, material and camera response all influence the observed color value significantly. These effects have been studied in the field of color imaging, leading to various algorithms for color constancy and for the computation of color invariants. At present these approaches mainly focus on global image features, and they are typically tested only on controlled high quality image databases.

Our research applies this work to local feature description. For shape characterization we rely on SIFT descriptors. The color descriptor will be used in combination with this so it does not need to contain spatial information but it does need to be robust to real-world photometric variations. This led us to represent color by local histograms of photometric invariants. Furthermore, to counter the instabilities of nonlinear color transformations we use weights based on an error-analysis to robustify the histograms.

We compared various different kinds of color histograms, each invariant with respect to different photometric variations, on a variety of tasks including matching, retrieval and classification. The reliability of our descriptors under geometric and photometric variations and image degradations was also established. Depending on the image set, the best results were sometimes obtained with color and sometimes with shape, but in all cases a combination of the two outperformed a purely shape-based approach [43].

A second approach [40] uses the *color flow model* to describe lighting changes between images depicting the same scene. Is based on two assumption: (a) the scene contains known objects and (b) the possible illumination changes are learned offline during a preprocessing step. These assumptions are restrictive, but they allow the influence of illumination changes on objects to be modeled very accurately. Although the model itself is linear, it is not restricted to linear lighting changes: the basis vectors that are chosen to describe the joint changes in color space are arbitrary and hence can capture complex nonlinear changes. This gives a powerful model while still allowing simple parameter estimation. During training we use aligned images of a static colored scene taken under different illuminations to learn the parameters of the model. The color flow model is then able to explain joint

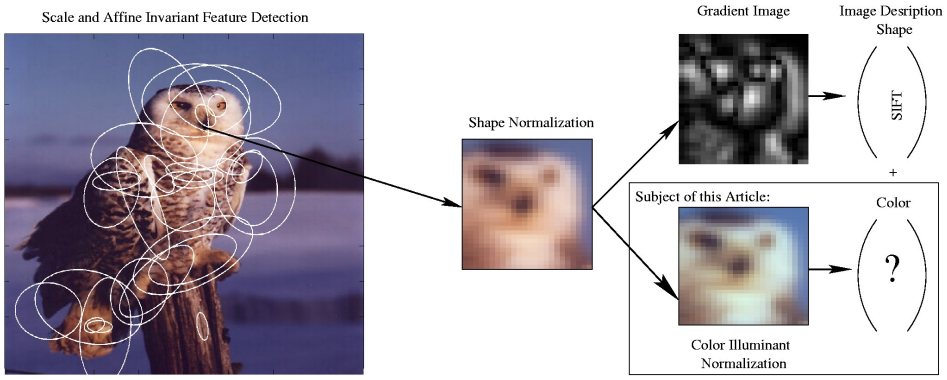


Figure 3: The local invariant descriptor method combines invariant local feature detection with robust local image description. The aim of this research is to enrich the local feature description with color information.

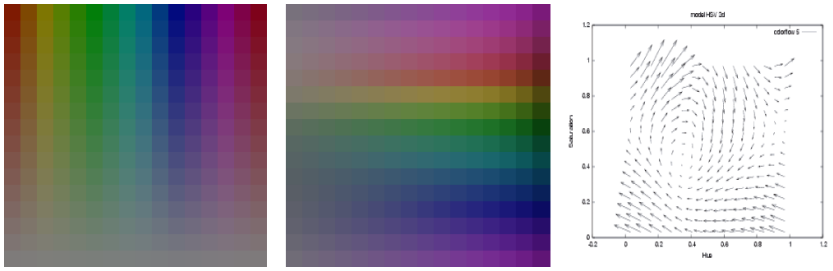


Figure 4: An illustration of the color flow model. Left: the reference color map (hue / saturation representation). Middle: the color map for a new illumination of the scene. Right: the transformation between the two maps – vectors represent individual color changes.

color changes between pairs of images, see figure 4.

We also take images of reference objects under normalized illumination. For these images we detect keypoints and store local descriptions of them in a database. When the method is used, an on-line normalization process extracts keypoints from the image and matches them to the keypoint database using local photometric signatures, thus allowing the color flow to be estimated robustly, after which all of the remaining image pixels can be normalized. This point correspondence process extends the original color flow method by relaxing the requirement for aligned images.

6.1.4 Indexing of visual descriptors

In order to perform image retrieval or visual object recognition based on SIFT and similar descriptors, it is often necessary to search a large database of previously seen descriptors for ones similar to those in the target image. This involves a search for neighbouring points or regions in a high dimensional feature space – a problem that is currently for practical purposes essentially linear in the database size, and hence slow for large databases. We developed an approximate method to speed up such

Figure 5: Examples of detections of spotted cats. The probability of a pixel being part of a spotted cat ranges from dark blue (low) through green (neutral) to dark red (high).

searches, based on the fact that the geometry of high-dimensional spheres allows strict bounding box bounds to be tightened quite a lot while still keeping most of the volume of the sphere. The method uses a KD-tree and inner and outer search radii, guaranteeing that all points within the inner radius, and most points within the outer radius, of the given point are recovered. This method is used in the demonstrators described in section 5.1.

6.1.5 Discriminative Regions for Semi-Supervised Object Class Localization

Labelling all of the individual pixels belonging to objects of a given class requires image features that can be extracted reliably and that cover the image densely. Salient regions defined by local interest points give reliable and repeatable object classification, but being sparse they can not label every pixel. In contrast, regions defined by unsupervised texture segmentation do cover the image, but although they can sometimes be reliable features, less textured objects can not be successfully represented. We have developed a method that combines the information from textured regions and salient local interest points to provide more robust pixelwise object localization. The method uses semi-supervised training data (only image-level labels) to automatically extract the most useful features for a given object class, be they region-based or interest point-based. It creates a region-level classifier based on the features chosen, combining the region's texture with the interest points found in and around the region. Some examples of the results are given in fig. 5.

6.1.6 Dense local descriptors

Another line of thought on image description is to keep the advantages of local descriptors, but to extract them at a dense set of patch locations and scales rather than sparsely at isolated keypoints. This greatly increases the amount of image information that can be encoded and reduces both per-patch computation and sampling bias as a separate feature detector is not needed. On the other hand, some of the local stability and invariance of the keypoint based methods is lost and many more patches need to be processed, most of which typically contain relatively little new information. In practice, simply replacing keypoint-based sampling with dense sampling does improve results for many image-level classification tasks, especially when using bag-of-features representations [36]. However we can do better by using an adpative clustering technique, see § 6.2.4.

6.2 Statistical modelling and machine learning for image analysis

Participants: Charles Bouveyron, Juliette Blanchet, Diane Larlus, Juho Kannala, Jakob Verbeek,

Cordelia Schmid, Bill Triggs, Frédéric Jurie, Florence Forbes [MISTIS], Stéphane Girard [LCM], Guillaume Bouchard [XEROX Research], Peter Carbonetto [University of British Columbia].

Keywords: semi-supervised learning, latent variable methods, dimensionality reduction, feature selection, graphical models, kernel methods, mixtures of experts.

6.2.1 High dimensional data analysis and clustering

The visual descriptors used in object recognition are usually high-dimensional but often lie in different low-dimensional subspaces of the original space. Global dimensionality reduction techniques are not useful in this case. We propose a method for discriminant analysis and clustering that finds the specific subspace and intrinsic dimension of each class. Our approach adapts a mixture of Gaussian framework to high-dimensional data: It determines class-specific subspaces and therefore limits the number of parameters that need to be estimated. The intrinsic dimension of each class is determined automatically using Cattell's scree test on eigenvalue sizes. The resulting approach for discriminant analysis [7, 29, 45] has shown good results for classification of visual descriptors, i.e. it outperforms linear SVMs (support vector machines).

The extension to clustering uses EM for parameter estimation and provides a robust clustering method in high-dimensional spaces that we call High Dimensional Data Clustering (HDDC). The number of parameters can be further limited by making additional model assumptions, for example assuming that classes are spherical in their subspaces or sharing some parameters between classes. We apply the method to probabilistic object recognition. Local scale-invariant features are clustered using HDDC, and the maximum likelihood based discriminative score for each cluster is determined using positive and negative examples of the category. By combining this information with the prior probabilities of the clusters we compute object probabilities for each visual descriptor. These are then used for object localization and image classification. Localization assumes that points with higher probabilities are more likely to belong to the object, and image classification is based on a per image score of the probabilities. Experiments on two recent object databases demonstrate the effectiveness of the clustering method for category-level localization and classification [31].

6.2.2 Markov models for the spatial organization of image descriptors

Our previous work on texture recognition [LSP03] introduced a method that simultaneously performs classification and segmentation. A generative model describes the distribution of the affine-invariant descriptors, along with co-occurrence statistics for nearby patches. At recognition time, initial probabilities computed from the generative model are refined using a relaxation step that incorporates co-occurrence statistics learned at modeling time.

Global co-occurrence statistics do not explicitly model the local dependencies between neighboring descriptors, and recognition results can be further improved by modeling these dependencies. Here we selected Markov Random Fields (MRF) as an appropriate statistically-based framework for this task. They provide parametric models whose parameters have a natural interpretation and they can be adjusted to incorporate a priori knowledge with respect to interaction strengths. Using them

[LSP03] S. LAZEBNIK, C. SCHMID, J. PONCE, "Affine-Invariant Local Descriptors and Neighborhood Statistics for Texture Recognition", *in : International Conference on Computer Vision, 1*, p. 649–655, 2003.

requires nontrivial parameter estimation. We use recent estimation procedures based on the mean field principle. The particularities of the application are the high-dimensionality of the feature vectors (typically more than 100 dimensional) and the irregularity of the sites at which they are observed. Very few practical optimization techniques are available for such tasks. They are usually very sensitive to initialization and to the parameters of the approach. By combining an MRF estimation procedure with a dimensionality reduction technique we show that recognition rates can be improved and that promising results can be obtained using a general statistical formalism. So far the work has focused on texture recognition [24, 25, 26], but future work will include other contexts such as object recognition.

6.2.3 A constrained learning approach to data association

As part of our collaboration with the University of British Columbia (UBC), P. Carbonetto from N. de Freitas' group applied UBC's approach for constrained semi-supervised learning by data association to the selection of discriminative local features [8]. Images are labeled as positive and negative, but individual descriptors are not labeled and may belong to the background even in positive images. Descriptors are labeled by constrained data association, where the constraints are on the number of positive descriptors in the image. The approach is based on a Bayesian classification model combined with an efficient Markov Chain Monte Carlo (MCMC) algorithm that simultaneously learns the unobserved labels and selects a sparse object class representation from the extracted high-dimensional descriptors. A generalized Gibbs sampler explores the space of labels that satisfy the constraints. Bayesian learning approximates the posterior distribution by integration over multiple hypotheses, a crucial ingredient for robust performance in noisy environments that also reduces the sensitivity to initialization. For practical MCMC exploration of the posterior modes, however, the posterior must be sufficiently peaked. This suggests the use of a Bayesian kernel model, as classical mixture models have numbers of modes that grow combinatorially with the number of components.

6.2.4 Adaptive Clustering for Visual Descriptors

Recently, "bag of features" methods based on comparing distributions of local patch descriptors have shown considerable promise for texture and category classification. They rely on vector quantization or some similar coding method to convert continuous high-dimensional visual patch descriptors to discrete codes that can then be histogrammed to characterize the image. The quantization codebook is a critical component of the approach.

Under dense coding, many patches are nearly empty and the distribution of patch descriptors in descriptor space is highly nonuniform with a single but very large central peak centred on the uniform patch. This disturbs a central element of the bag-of-features approach – the construction of the codebook that is used to vector quantize descriptors. K-means based codebooks are usually used in the sparse case, but in the dense one their code centres cluster densely around the central peak, thus starving the rest of the space of codes and giving a relatively poor coding. Better adapted codebooks can further improve the results of the dense method. We have developed an efficient fixed-radius-based clustering method that corrects the starvation effect, giving significantly better codebooks and greatly improving the results of dense patch based classification. It provides codebooks with better distributed centres, thus ensuring that regions with comparatively low probabilities are still coded well.

At each step our method discovers new clusters by subsampling the data and running K-medians

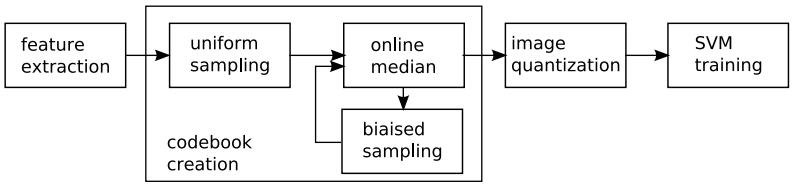
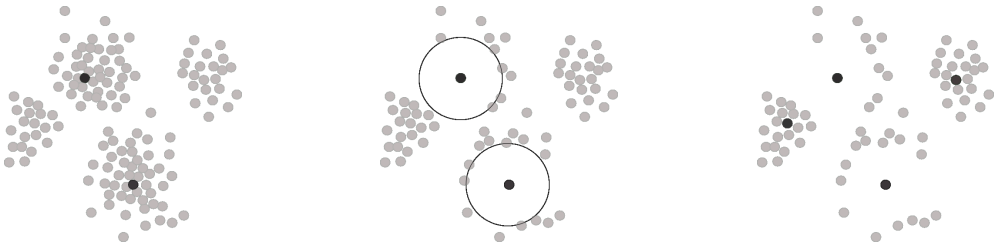


Figure 6: Adaptive clustering method. Outline of the learning steps. See text for details.

(see fig. 6). It iterates until the required total number of clusters has been found. The influence of highly populated regions is decreased by using biased subsampling to force new centres to be well separated from previously found ones – points that lie closer than a given *influence radius* to a previous centre are excluded from the sample. The centre selection algorithm is Mettu & Plaxton's *online median*, which is based on the *facility location* problem. It chooses the centers one by one but it is run several times so that a few new centers are added at each iteration. Fig. 7 illustrates the process. Suppose that we discovered 2 new centers in the previous step – the two black points in fig. 7(a). Points within the influence radius of these are excluded from further sampling, fig. 7(b), and the next phase of K-medians is run, fig 7(c). The key parameters of the method are the influence radius and the total number of clusters. Distances are measured by Euclidean distance in SIFT space.



(a) (b) (c)

Figure 7: Biased sampling in the adaptive clusterer.

Coding the feature patches using the final codebook gives a very effective *bag-of-features* representation for image classification [36], for example winning several categories of the PASCAL Network of Excellence Visual Recognition Challenge.

6.3 Visual recognition and content analysis

Participants: Ankur Agarwal, Navneet Dalal, Gyuri Dorko, Diane Larlus, Marcin Marszalek, Eric Nowak, Jianguo Zhang, Cordelia Schmid, Bill Triggs, Frederic Jurie, Roger Mohr, Svetlana Lazebnik [UIUC], Jean Ponce [UIUC], Guillaume Bouchard [XEROX Research].

Keywords: visual models, object class recognition and detection.

6.3.1 Recognition of texture and object classes – local features and kernels

We performed a large-scale evaluation of a classification framework in which images are represented as distributions (signatures or histograms) of features extracted from a sparse set of keypoint locations [47, 20, 9]. Classification is performed using Support Vector Machines with kernels based on two effective metrics for comparing distributions – the Earth Mover's Distance and the chi-square distance. We first evaluate the performance of our approach with different keypoint detectors and descriptors, as well as different kernels and classifiers. This has shown that combining multiple detectors and descriptors typically gives better results than the best individual detector/descriptor channel, and including too much invariance in the local features damages the results. So for the best performance one should incorporate several complementary features and select the invariance properties that provide just the level of invariance needed for the application, but no more.

We evaluated our methods against several state-of-the-art algorithms on four texture and five object databases. In most cases our method met or exceeded the best previously reported results, see table 6.3.1 for a comparison with the best reported results on object categories. Note that the approach won all four visual object classification tasks of test2 proposed during the European Union Network of Excellence PASCAL's Visual Recognition Challenge. The power of bag-of-keypoint representations for texture classification is perhaps not surprising given that the images have uniform statistical properties and little or no global spatial organization and clutter. But it is not obvious *a priori* that such representations also obtain good results for object category classification as they ignore spatial relations and do not separate foreground from background features. In the longer term, successful category-level object recognition and localization is likely to require more sophisticated models that capture the 3D shape of real-world object categories as well as their appearance. In the development of such models and in the collection of new datasets, simpler bag-of-keypoints methods can serve as effective baselines and calibration tools.

	Xerox7	Caltech6	Graz	Pascal		Caltech101
				test set1	test set2	
ours	**94.3**	**97.9**	**90.0**	92.8	**74.3**	**53.9**
others	82.0	96.6	83.7	**94.6**	70.5	43

Table 1: Comparison with the best reported results on several object datasets.

Many methods including bags-of-keypoints typically make classification decisions about the image as a whole, thus potentially using both foreground features and background "context". Whether you consider context to be a valid cue or a cheat, in many datasets it is well enough correlated with the foreground to be a potential source of information. For example, cars are frequently seen on a road or in a parking lot, while faces tend to appear against indoor backgrounds.

To study this effect we used the PASCAL Visual Recognition Challenge dataset for which ground-truth object localization information is available, separating the foreground and background features and evaluating each separately. Particularly for the easier datasets, the experiments revealed that the backgrounds alone do in fact contain a considerable amount of discriminative information for the foreground category. However including both foreground and background features *never* improves performance relative to foreground features alone – it is the features on the objects themselves that play the key role in recognition in our distribution based approach. In fact, we show that it is dangerous

to train recognition systems on datasets with monotonous or highly correlated backgrounds – such systems tend to overfit to the background, so they generalize poorly to more complex test sets.

6.3.2 Recognition and Localization of Object Classes – Feature Selection

Our earlier work [DS03] on discriminative feature selection was extended to weakly supervised scenarios and a wider range of object categories [46]. We have also integrated our approach into an object class localization framework [LS04] whose goal is to determine bounding rectangles around each instance of the object category. In this case training is done in a supervised fashion in which all objects are given with their corresponding rectangles. Our method first learns a vocabulary from the scale-invariant features of the training set using expectation-maximization (EM) to estimate a Gaussian Mixture Model with a diagonal covariance matrix. We then assign a rank to each cluster based on its discriminative power [46] and learn the spatial distribution of the object positions and scales for each cluster. For each training image, we assign all descriptors inside the object's bounding rectangle to its cluster using MAP, and record the center and scale of the rectangle with respect to the assigned cluster. This step is equivalent to [LS04] except that we collect the width and height separately and we do not use the figure-ground segmentation of the object. The output of our training is a list of clusters with the following properties: (i) a mean and variance representing the appearance distribution of the cluster; (ii) a normalized (probabilistic) score for its discriminative power; and (iii) a spatial distribution over the object positions and scales.

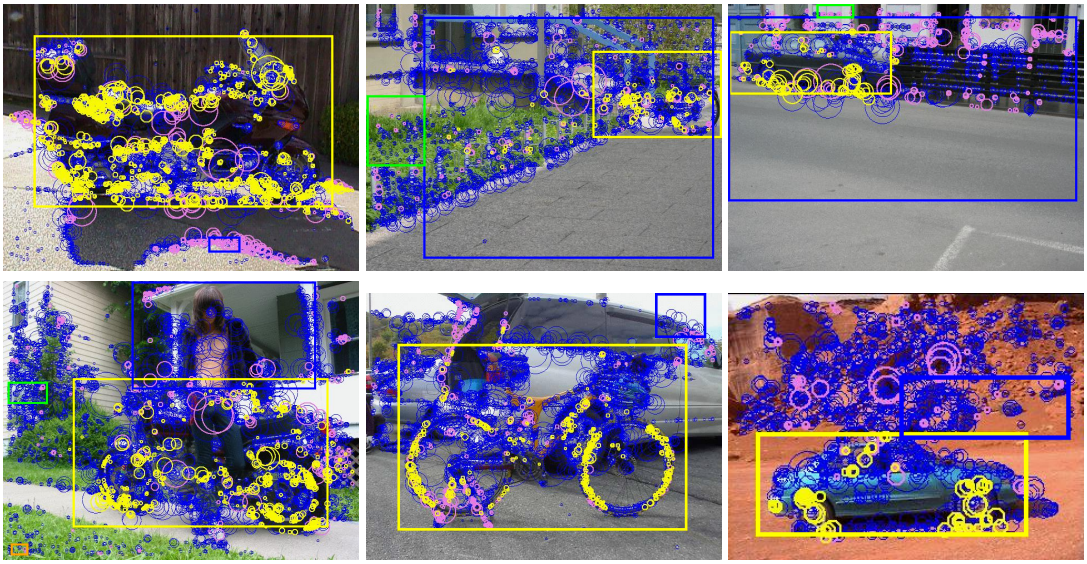
Object localization in a test image is similar to the initial hypothesis generation used during training, except that we incorporate a discriminative element into the voting scheme: to allow better confidence estimates for the different hypotheses, only the n most discriminative clusters participate in the voting, and their probabilistic scores are integrated into the voting. The extracted scale-invariant features of the test image are assigned to the closest cluster by appearance (MAP). Then, the chosen clusters vote for possible object locations and scales (4D space). Fig. 8 shows some examples of detections on test images from the PASCAL dataset.

6.3.3 Hyperfeatures – multilevel local coding for visual recognition

Several recent visual coding methods for image classification and object recognition are based on histograms of appearance descriptors evaluated on local image patches. Such descriptors can be highly discriminant and they have good resistance to local occlusions and to geometric and photometric variations, but they are not able to exploit spatial co-occurrence statistics at scales larger than their local input patches. We have developed a new multilevel visual representation, 'hyperfeatures', that is designed to remedy this. The starting point is the familiar notion that to detect object parts, in practice it often suffices to detect co-occurrences of more local object fragments – a process that can be formalized as comparison (*e.g.* vector quantization) of image patches against a codebook of known fragments, followed by local aggregation of the resulting codebook membership vectors to detect co-occurrences.

[DS03] G. DORKO, C. SCHMID, "Selection of Scale-Invariant Parts for Object Class Recognition", *in : International Conference on Computer Vision, 1*, p. 634–640, 2003.

[LS04] B. LEIBE, B. SCHIELE, "Scale Invariant Object Categorization Using a Scale-Adaptive Mean-Shift Search", *in : Proceedings of the DAGM'04 Annual Pattern Recognition Symposium, Tuebingen, Germany, 3175*, Springer LNCS, p. 145–153, August 2004.



Motorbikes Bicycles Cars

Figure 8: Examples of detections on the PASCAL dataset, test set 1 (top) and test set 2 (bottom). The blue points were eliminated by feature selection, while yellow points vote for the best solution (the yellow rectangle). Non-yellow rectangles are alternative detections with lower confidence.

This process converts local collections of image descriptor vectors into slightly less local histogram vectors – higher-level but spatially coarser descriptors. As the output is again a local descriptor vector, the process can be iterated, and doing so captures and codes ever larger assemblies of object parts and increasingly abstract or 'semantic' image properties. We have studied the performance of hyperfeatures extensions of several different image coding methods including clustering based Vector Quantization and Gaussian Mixtures, finding that the latter consistently outperform the former, and that adding a stage of Latent Dirichlet Allocation – a probabilistic "topic distillation" model that has recently been developed in the statistical text community – improves the results further. The resulting high-level features provide improved performance in several object image and texture image classification tasks. We are currently in the process of developing the method for object localization. Some results demonstrating this are shown in figure 9. The work is described in [44].

6.3.4 Semi-local parts models

We use characteristic patterns formed from scale or affine-invariant patches linked by semi-local spatial relations to describe salient object parts [38]. These invariant semi-local parts are geometrically stable configurations of multiple invariant regions, found by the Laplacian keypoint detector or by our shape detector [JS04]. The parts are more distinctive than individual features, and their locality makes the method suitable for modeling a wide range of 3D transformations, including viewpoint changes and non-rigid deformations. They are learned using the idea that direct search for visual correspondence is

[JS04] F. JURIE, C. SCHMID, "Scale-invariant shape features for recognition of object categories", *in : IEEE Conference on Computer Vision and Pattern Recognition, II*, p. 90–96, Washington DC, USA, 2004.

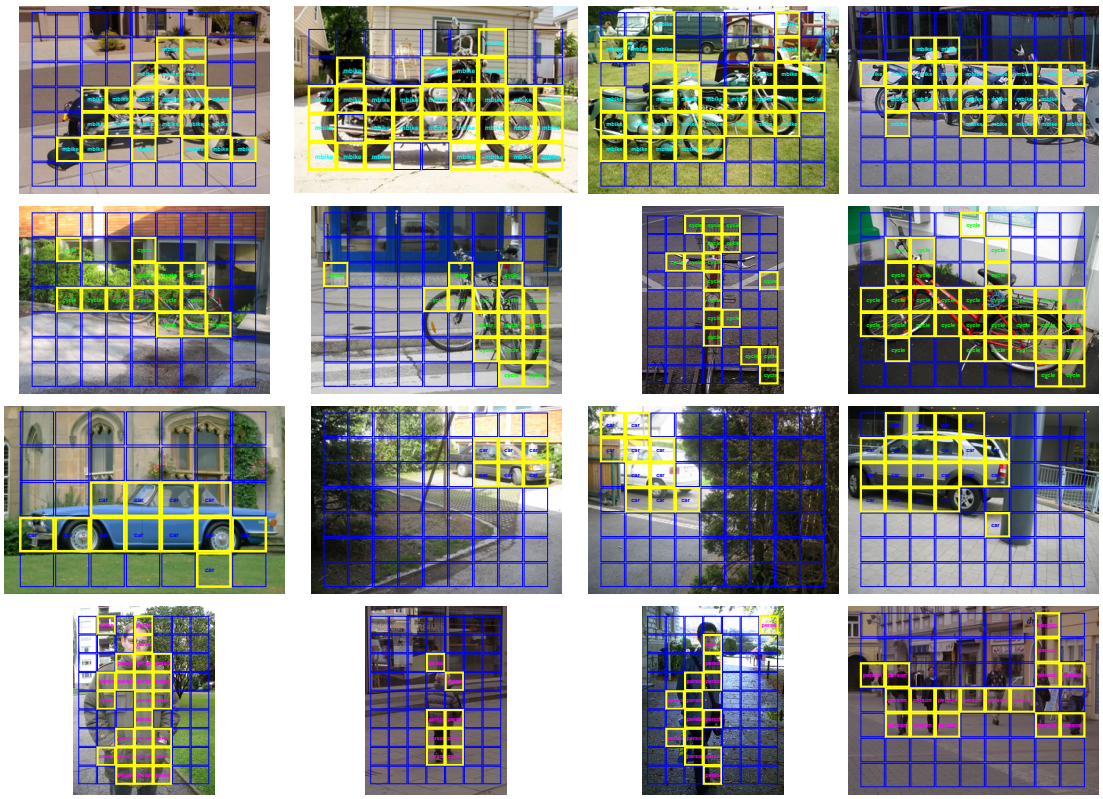


Figure 9: Object localization on 4 object categories from the PASCAL object recognition challenge dataset, based on classifying each local image region using its hyperfeatures. Each row shows examples of results using one of the four independent classifiers, each being trained to classify foreground regions of its own class against the combined set of all other regions – background regions and foregrounds from other classes. Currently, each region is classified independently – no spatial smoothness constraint is enforced.

the key to successful recognition, see fig. 10(a) for the most reliable, i.e. highest-scoring, parts of each category. A discriminative maximum entropy framework is used to learn the posterior distribution of the class label given the occurrences of parts in the training set. Experimental tests have demonstrated the effectiveness of the framework for visual classification, see figs. 10 and 11.

6.3.5 Hierarchy of parts models for object category recognition

The robust encoding of natural shape and geometry is one of the main challenges of visual recognition. Above we have seen several approaches to this based on local features models: bag of features, which use large sets of local features without any explicit representation of geometry; semi-local parts, which combine small sets of local features in relatively rigid arrangements; and hyperfeature hierarchies which code simple co-occurrence of local features but do so at several levels. An alternative is to subject the local features to a loose but still geometric spatial model that quantifies their relative

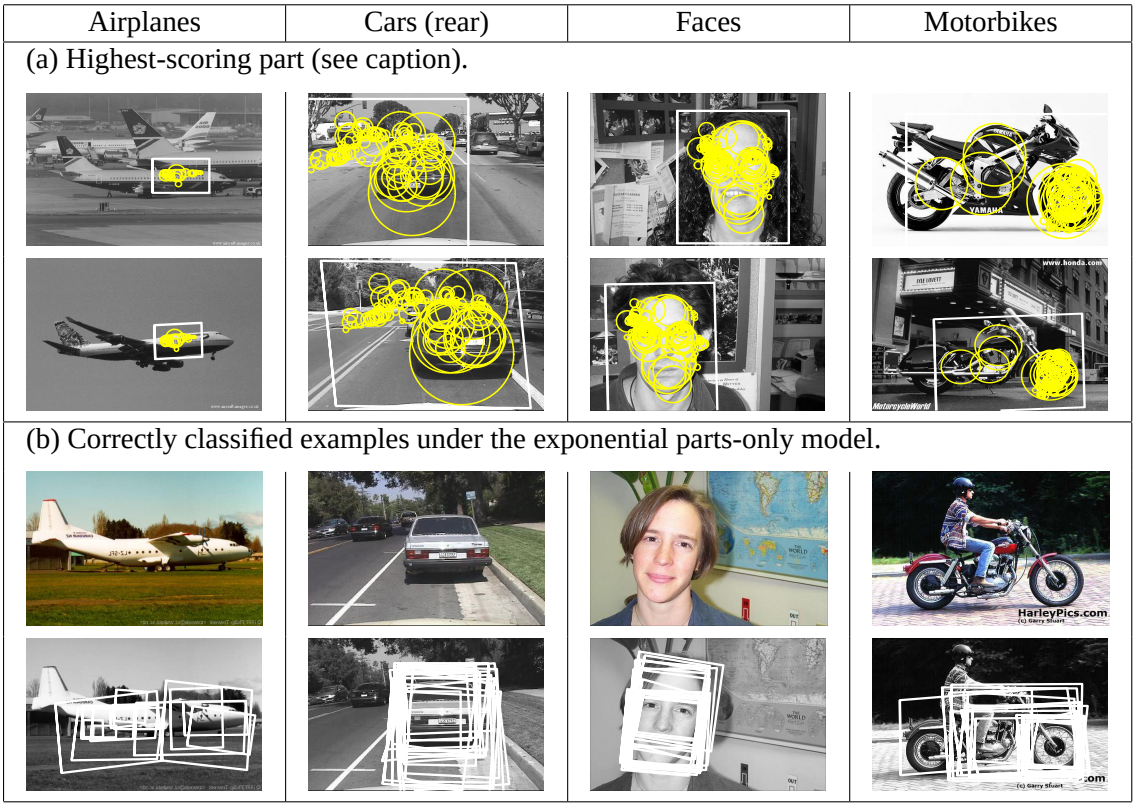


Figure 10: Semi-local parts – results on the Caltech dataset. (a) The highest-scoring part for each class. The two training images that were originally matched to obtain the part are shown, with the matched regions (yellow circles) superimposed. The aligning transformation between the two groups of matches is indicated by the bounding boxes: the axis-aligned box in the top image is mapped onto the parallelogram in the bottom one. (b) An example of a correctly classified image for each class. The original images and the transformed bounding boxes of the parts matched are shown. The localization is poor for airplanes but very good for faces. For motorbikes, the front wheel is particularly salient.

geometry more globally (at a whole-object level). We have developed a generative probabilistic model of this kind that codes the geometry and appearance of a visual object category as a loose hierarchy of parts, with probabilistic spatial relations linking parts to subparts, soft assignment of subparts to parts, and scale invariant keypoint based local features at the lowest level of the hierarchy. The framework efficiently handles models containing hundreds of redundant local feature classes, such as those returned by current keypoint detectors. The resulting degree of redundancy allows the method to outperform constellation style models, despite their stronger spatial models. Models are instantiated by robust bottom-up voting over a hierarchy of location-scale pyramids (one for each part), and optimized by Expectation-Maximization. Training is rapid, and there is no need for object positions to be marked in the training images. Experiments on several popular datasets show the method's ability to capture complex natural object classes. Figure 12 sketches the structure of the model, and figure 13 shows examples of how it adapts to changes of viewpoints and the intra-class variations of a generic visual

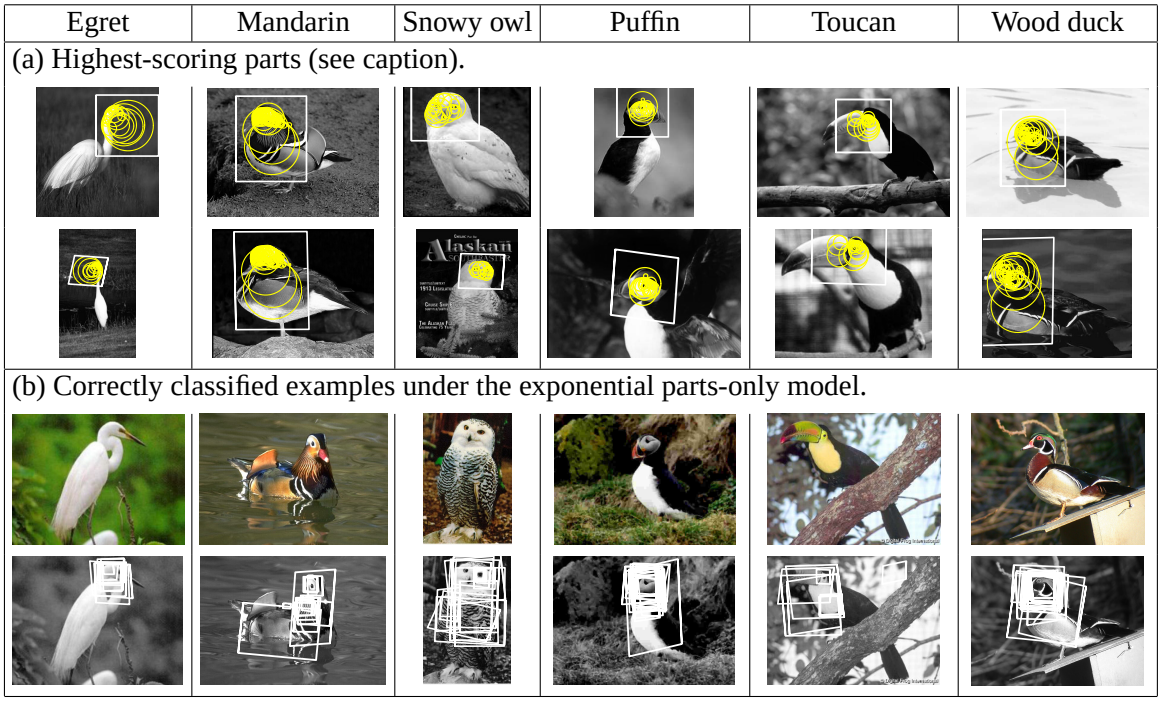


Figure 11: Semi-local parts – results on Birds database. (a) First and second rows: the highest-scoring part for each class superimposed on the two original training images. (b) A successfully classified image for each class, showing the original test image (top), and the image with the bounding boxes of all of the detected class part instances superimposed (bottom). Notice that localization is fairly good for mandarin and wood ducks (the head is the most distinctive feature). The owl parts are more prone to false positives but they do still capture the salient characteristics of the class: the head, the eye, and the pattern of the feathers on the breast and wings.

object class.

6.3.6 Classification into class hierarchies based on local parts

The above models embody one-of-n classification into flat sets of visual classes. It is also important to achieve more structured kinds of labelling such as categorization of visual objects into hierarchically organized classes. We have built a model of this kind based on bag of features descriptors, that can handle visually similar classes despite high within-class variability [41]. As usual, the local parts approach provides robustness to pose, occlusions, illumination and shape variations. For visually similar classes there are typically only a few parts that are discriminative enough to separate the classes reliably, so the selection of appropriate input parts is critical. The method achieves this using variable selection techniques from machine learning and organizes the resulting parts in a tree structure. The parts at the top of the tree provide coarse inter-category discrimination, while those at the bottom capture category specific details (see fig. 14). The hierarchical structure of the input classes is exploited to select the best parts for each level of the tree. A particular focus is the trade-off between computation

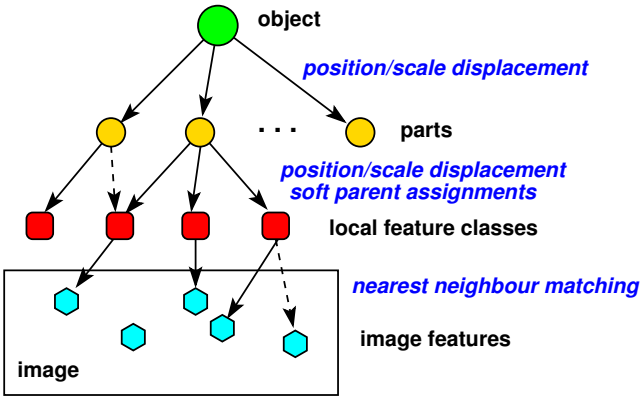


Figure 12: Our hierarchy of parts model for visual object recognition is a tree-structured hierarchy of parts and subparts with the complete object at the root and individual scale-invariant local feature classes at the leaves. A probability distribution over geometric transformations between each subpart and its parent quantifies the subpart's relative position and uncertainty. The parent attributions for each sub-part are also uncertain and are learned during training. During model instantiation the leaf parts are coupled to the nearest observed image features in position and appearance via a robust observation model that effectively ignores unattributed features.

time (number of parts used) and recognition rate. We have tested the method on visible images and on infrared images of military vehicles acquired by surveillance cameras. On the infrared dataset, for the same run time, the accuracy of our hierarchical method is 12% better than a standard one-versus-one SVM.

6.4 Human detection and activity analysis

Participants: Ankur Agarwal, Navneet Dalal, Bill Triggs, Cordelia Schmid, Cristian Sminchisescu [Toronto].

Keywords: human detection, human body pose and motion, activity analysis.

6.4.1 Human detection – histogram of oriented gradient descriptors

As part of our ongoing work on human detection we have developed [33] a robust feature set for visual object recognition in general, and for "pedestrian detection" (the detection of upright, full visible, standing or walking people) in particular. The new "Histogram of Oriented Gradient (HOG)" feature set is inspired by the success of SIFT descriptors [Low04a] for local feature based recognition, but here, rather than being sampled only sparsely at salient local feature points, blocks of well-normalized gradient orientation histograms are used in a dense grid, at a uniform scale and without rotation normalization. This provides a particularly robust and discriminant appearance based descriptor suitable

[Low04a] D. LOWE, "Distinctive Image Features from Scale-invariant Keypoints", *International Journal of Computer Vision 60*, 2, 2004, p. 91–110.

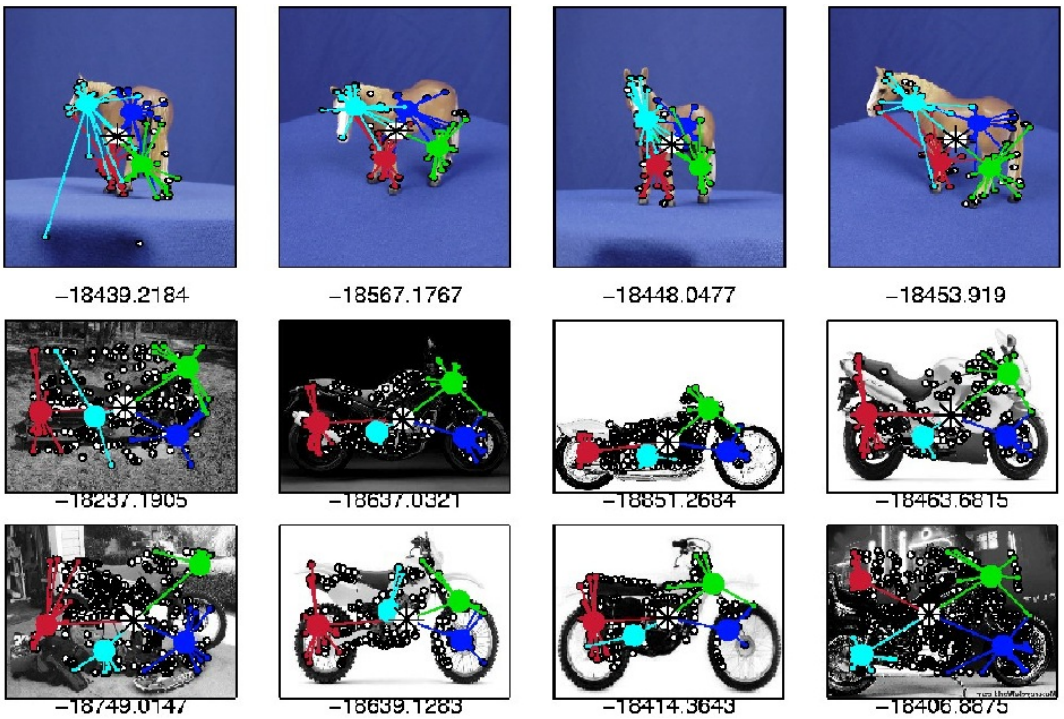


Figure 13: Some examples of our hierarchy of parts model in action. The first row illustrates that the model has good resistance to viewpoint changes owing to its loose hierarchical structure, the second that it can adapt to class variations, here different types of motorcycle.

for visual classes that have a reasonable degree of spatial consistency across examples. The current overall detector uses a monolithic (non-parts-based) linear Support Vector Machine as a classifier in the HOG window, scanning this densely across the test image at multiple scales followed by a mean-shift based space-scale peak location method to recover multi-scale detections.

We performed a detailed experimental evaluation on existing data sets and on a more difficult new 2400 image dataset that was created for the purpose. The HOG features significantly outperform existing ones including shape contexts and various types of wavelets, giving 1-2 orders of magnitude less false positives than the best previous detectors. The study also shows that small derivative scale, fine orientation sampling, moderately coarse spatial sampling, good local normalization and significant overlap between descriptor windows all significantly improve the performance. Fig. 15 summarizes the performance of various different descriptors, and fig. 16 shows which features within the detector window are most important for the detection process.

Overall the detector has proven very popular, with at least 6 other industrial and academic research groups reimplementing it for human detection and many downloads of both the detector binaries and the new human detection dataset (http://pascal.inrialpes.fr/soft/olt).

With different training sets, the same approach was used in the PASCAL Network of Excellence Visual Recognition Challenge, winning several of the object detection categories including humans and cars [9].

Recently we have extended the method to include a differential optical flow channel to provide

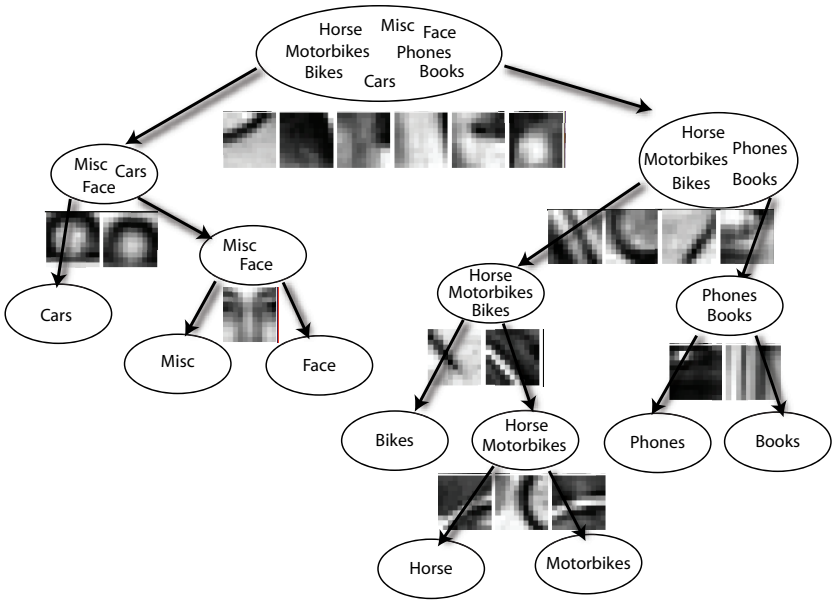


Figure 14: Object classification into class hierarchies. At each level of the classifier, most discriminant features are selected. The hierarchical organization of the classes is automatically computed.

improved discrimination for human detection in films and videos [32]. The method allows stationary or moving subjects, non-rigid or moving backgrounds and moving cameras. It provides a further order of magnitude reduction in false positive rate.

6.4.2 3D human pose and motion from monocular images – model based approach

Over the past several years we have investigated several approaches to the problem of "motion capture" (the estimation of articulated human pose and movement) from monocular images and image sequences. Previously we took a model based approach to this. Owing to the large number of criss-crossing kinematic solutions in this problem, this approach necessitated the development of advanced methods for finding nearby local minima of the complex high-dimensional model-image matching cost function. This year has seen the publication of a journal paper on modified local-descent optimization methods for approaching this issue [17]. The methods used are generic and may be of interest in many other optimization problems with smooth non-convex cost functions with moderately large numbers of local minima.

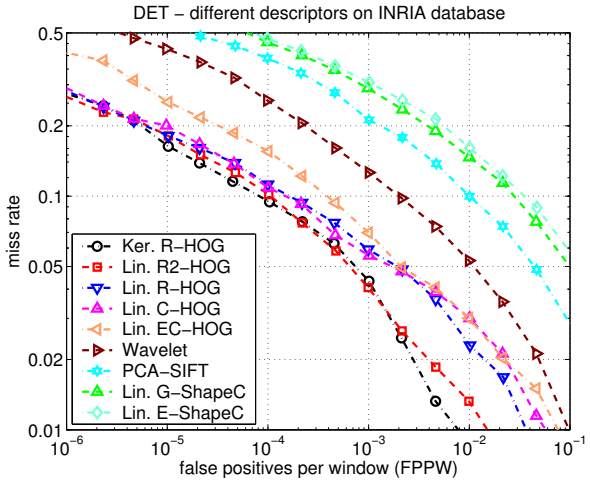
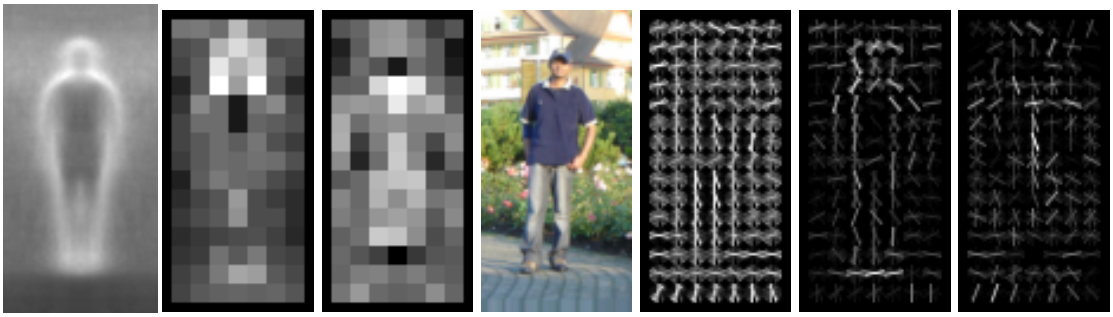


Figure 15: Human detection. A summary of the performance of the different classes of descriptors tested, on our new test database using linear SVM and the same training paradigm in all cases. The Histogram of Oriented Gradient (HOG) features have false positive rates 1-2 orders of magnitude lower than the best wavelet features. Replacing the linear SVM with a Gaussian kernel one further increases the descriptor performance by about 3% at 10^{-4} false positives per window, at the cost of much higher running time.



(a) (b) (c) (d) (e) (f) (g)

Figure 16: The HOG based human detectors cue mainly on silhouette contours, especially ones around the head, shoulders and feet. (a) The average gradient image over the training examples – despite changes in pose of the subjects, there is a good deal of consistency. (b,c) Each "pixel" shows the maximum positive (b) or negative (c) SVM weight in the block centred on the pixel. The most active positive blocks are those centred on the image background just *outside* the contour, i.e. contour gradients are normalized w.r.t. the background not w.r.t. the interior of the figure. The most active negative blocks are inside the figure, cancelling out, e.g. responses from long vertical edges like trees and poles that happen to be aligned with the detector legs. (d) A test image. (e) Its computed HOG descriptor. (f,g) The descriptor reweighted by respectively the positive and the negative SVM weights. The use of the head, shoulders and foot contour and the cancellation of internal edges is again evident.

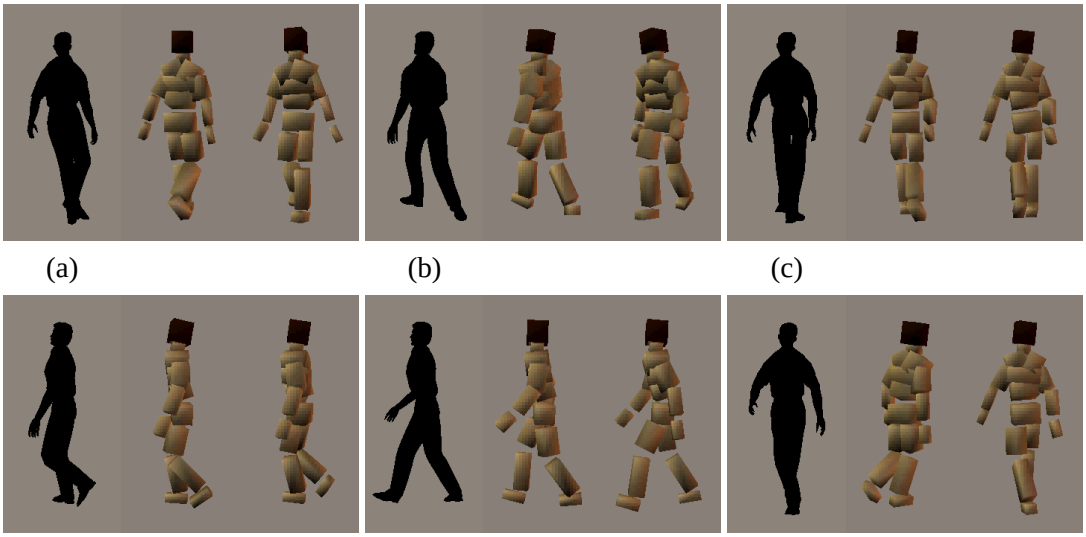


Figure 17: Mixture of regressors approach to monocular pose recovery. Each figure shows a silhouette and the two 3D poses that are the most likely for the silhouette based on our multivalued regression model. The model typically captures the two most intuitive alternatives, illustrating cases of forward-backward ambiguity (a,b), kinematic flipping of the legs (c) and interchanging labels between the two legs (d,e). (f) shows an example where the model's most probable solution is clearly incorrect, but feasible solutions are obtained in the other modes.

6.4.3 3D human pose and motion from monocular images – learning based approach

As an alternative to the model based approach, we began to develop example (learning) based approaches to the monocular human motion capture problem last year [AT04a,AT04b], and we have continued this work this year. The essential advantage of example-based formulations is that they capture poses and motions that are *typical* whereas the geometric model based approach is much less constrained – it permits any pose that is *possible*, no matter how unlikely. The main disadvantages of example-based methods are the need for large amounts of training data and the effective restriction to poses that are not too dissimilar to ones that have been seen during training. Rather than relying solely on examples, we have adopted sparse regression based approaches to learning, as regression allows regularization, thus improving generalization, and sparseness reduces the number of active model parameters or training examples that need to be accessed at run time.

3D human motion from silhouettes. Last year we developed several variants on learning based approaches to monocular human motion capture based on sparse kernel regression over a robust representation of silhouette geometry based on vector quantization coding of shape context descriptors.

[AT04a] A. AGARWAL, B. TRIGGS, "3D Human Pose from Silhouettes by Relevance Vector Regression", *in : IEEE Conference on Computer Vision and Pattern Recognition*, p. II 882–888, Washington DC, USA, June 2004, `http://lear.inrialpes.fr/pubs/2004/AT04`.

[AT04b] A. AGARWAL, B. TRIGGS, "Learning to Track 3D Human Motion from Silhouettes", *in : International Conference on Machine Learning, Banff, Canada*, p. 9–16, July 2004, `http://lear.inrialpes.fr/pubs/2004/AT04b`.

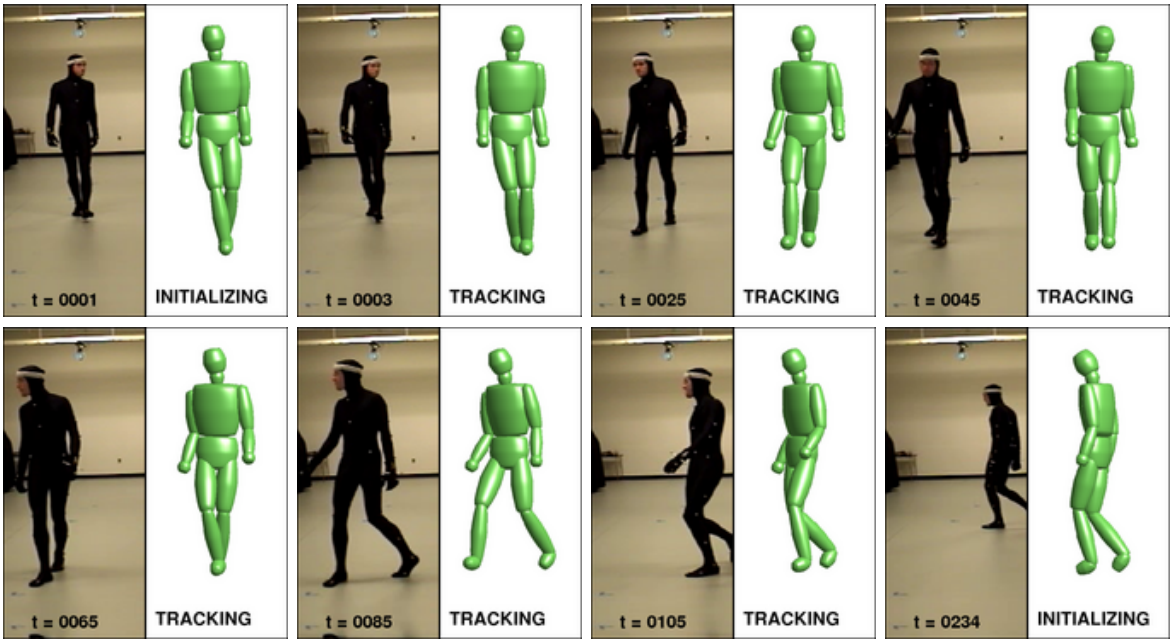


Figure 18: Multiple hypothesis tracking of a person over 500 frames using the mixture of regressors method. (Sequence taken from http://mocap.cs.cmu.edu). Our method provides direct probabilistic pose estimation from a single image, thus allowing automatic initialization and re-initialization. Maintaining multiple track hypotheses allows the tracker to recover from possibly inaccurate initializations, tracking stably through regions in which the person is not observed.

This approach works reasonably well most of the time, but when used for single image recovery (outside of a temporal tracking framework), it is subject to 'glitches' in which the incorrect pose is reconstructed. These occur because the silhouette based representation itself is ambiguous: it is often impossible even for a human to decide which leg or arm is foremost, whether the silhouette was seen from the back or the front, etc. This year we have developed a method that works around this by returning a set of possible reconstructions with associated probabilities [22]. Technically, the approach centres around a mixture of probabilistic regressors. It uses local clustering in the input and output spaces to identify regions of multi-valuedness in the mapping from silhouette to 3D pose, and uses these to initialize the fitting of a mixture of experts on the input manifold. The resulting mixture of nonlinear regressors predicts a (small) set of probable pose solutions. These can then be used, e.g., in a particle filter framework to provide a robust tracker that is capable of recovering automatically from mistracking events. Examples of silhouette ambiguities are shown in fig. 17 and some tracking results are shown in fig. 18. A forthcoming journal paper details both this and our previous work on learning based motion capture [5].

Monocular human pose from cluttered images. The above regression based methods for monocular human motion capture give very promising results, but they are based on silhouettes. They thus require a relatively clean background, or at least a stationary one on which background subtraction can be used. This year, we have begun work on extending the learning based approach to unsegmented

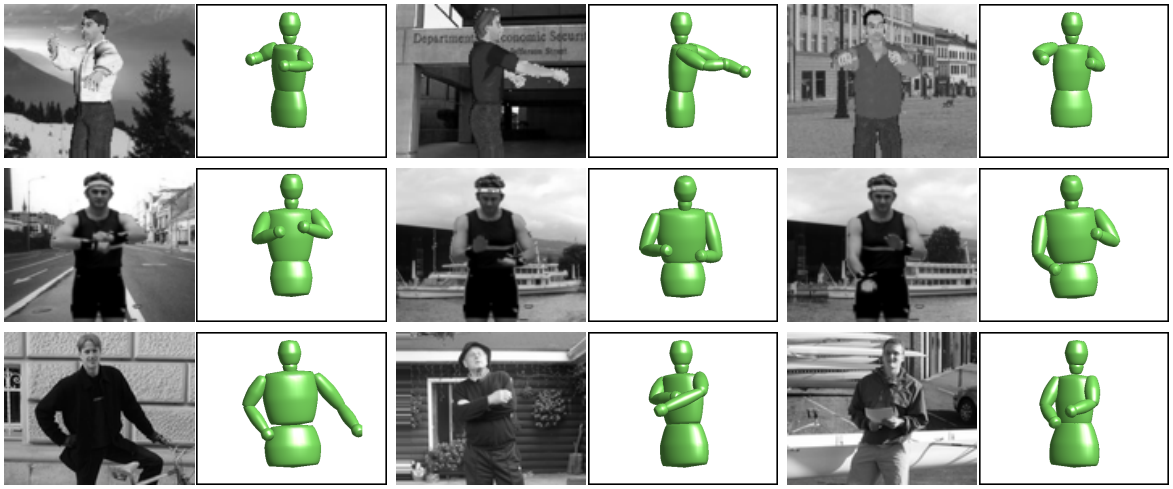


Figure 19: Pose reconstructions from our cluttered background method, on artificially rendered and real images with complex backgrounds. Top row: unseen test images with a random background image and an artifical human in a random pose synthesized using POSER. Middle row: a test sequence of a real person from our motion capture dataset, again with artifical backgrounds to make the problem more difficult. The training set included similar gestures made by another person. Bottom row: unseen real images obtained using Google. The method is purely appearance based – no segmentation of the subject is needed.

images with complex, cluttered backgrounds [23]. In keeping with the philosophy of a bottom-up example-based approach without explicit body modelling, we use learning to obtain a set of image features that allows pose to be regressed directly from unsegmented images. Currently the method supposes that the human subject has already been approximately located in the image. It takes an image window containing him, evaluates a dense grid of local gradient orientation histograms in the window (as used in the above human detector), reduces each descriptor with respect to a local basis learned using non-negative matrix factorization, and uses the resulting basis coefficients for pose regression as in the previous methods. Using non-negative factorization allows the method to selectively encode human-like features and to suppress much of the background clutter. For example it can key on patterns, such as shoulder contours and bent elbows, that are characteristic of humans and that carry significant pose information. The system was trained on a database of images with poses obtained using conventional multi-camera motion capture. It has comparable accuracy to other existing example-based methods, but unlike them it continues to work on unsegmented images with complex natural backgrounds. Fig. 19 shows some examples.

7 Contracts and Grants with Industry

7.1 Bertin Technologies

Participants: Eric Nowak, Frederic Jurie, Roger Mohr.

The collaboration with Bertin Technologies centres on developing algorithms for detecting and recognizing objects in unmanned infra-red information systems. Typical applications are outdoors defense systems in which hidden cameras are left to detect the presence of military vehicles. The main challenges are the relatively poor image resolution, the changeable appearance of objects due to global and local temperature changes, and the potentially large number of nested object categories. The project funds the CIFRE grant for Eric Nowak's PhD thesis, which started in March 2004. We plan to extend the collaboration in 2006. Bertin Technologies also participates in our Techno-Vision project ROBIN (see paragraph 8.1.2).

7.2 MBDA Aerospatiale

Participants: Julien Bohne, Frederic Jurie, Cordelia Schmid.

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

7.3 THALES Optronics

Participants: Frederic Jurie, Diane Larlus.

In 2004-2005 we collaborated with THALES Optronics. The aim was to develop algorithms for detecting objects in aerial images, with a focus on selecting reliable visual parts by clustering large sets of features. We plan to continue this collaboration in 2006.

7.4 EADS Fondation

Participants: Vittorio Ferrari, Frederic Jurie, Cordelia Schmid.

The postdoctoral scholarship of Vittorio Ferrari is financed by the EADS Foundation. The project started in November 2005 and will explore image contours as an alternative representation for visual class recognition. In contrast to systems based on local invariant textured patches, this will allow classes that are mostly defined by their shape to be recognized, such as bottles, mugs, or horses.

7.5 Siemens Corporate Research

Participants: Navneet Dalal.

Following Navneet Dalal's successful internship at Siemens' Princeton lab in 2004, they financed him for several months in 2005 to participate in the production of a book on the Mean Shift method.

8 Other Grants and Activities

8.1 National Projects

8.1.1 Ministry grant MoViStaR

Participants: Charles Bouveyron, Juliette Blanchet, Jianguo Zhang, Cordelia Schmid, Bill Triggs.

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

8.1.2 Techno-Vision project ROBIN

Participants: Benjamin Ninassi, Frederic Jurie, Roger Mohr, Cordelia Schmid.

We lead the national Techno-Vision project ROBIN, which started in January 2005 for two years. The aim is to quantify and consolidate progress in visual object recognition by developing ground truthed datasets and performance metrics to improve the evaluation of object recognition algorithms, and by running a national competition in this area. The project is funded partly by the French Ministry of Defense, the French Ministry of Research (Techno-Vision funds) and by several companies and research centers (Bertin Technologies, Cybernetix, DGA, EADS, INRIA, ONERA, MBDA, SAGEM, THALES and 35 public laboratories). It will cover multi-class object detection, generic object detection, generic object recognition, and image categorization. During the first year (2005/2006) the project produces datasets and metadata, while the second year will be devoted to selecting the test images and the benchmarking procedure as well as organizing the competition.

8.2 European Projects and Grants

8.2.1 FP5 Project LAVA

Participants: Guillaume Bouchard, Gyuri Dorko, Michael Sdika, Matthijs Douze, Ankur Agarwal, Peter Carbonetto, Cordelia Schmid, Bill Triggs, Roger Mohr.

Learning for Adaptable Visual Assistants (LAVA) was a 5th framework RTD project running from May 2002 to April 2005. It developed advanced machine learning based computer vision techniques for understanding everyday scenes, in particular for applications suited for embedding in camera-equipped electronic devices such as personal assistants and portable telephones. It was an interdisciplinary project, involving teams working on machine learning, computer vision, and cognitive modeling and data fusion. The coordinator was Xerox Research Centre Europe (XRCE, Grenoble, France), and the other partners were: LEAR; VISTA (IRISA-INRIA, Rennes, France); Royal Holloway College and the University of Southampton (Egham and Southampton, U.K.); Lund University (Lund, Sweden); Graz Technical University and the University of Leoben (Graz and Leoben, Austria); the Institut

Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP, Martigny, Switzerland); and the Australian National University (ANU, Canberra, Australia). In total it involved 51 person-years of research effort, for a total budget of 4.3 MEu, including 2.4 MEu of European Union support. LEAR worked mainly on the development of image descriptors and robust visual correspondence, on the interface between vision and learning, and on semi-supervised learning.

8.2.2 FP6 Integrated Project aceMedia

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

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

8.2.3 FP6 Project CLASS

Participants: Bill Triggs, Cordelia Schmid, Frederic Jurie, Jakob Verbeek.

CLASS (Cognitive-Level Annotation using latent Statistical Structure) is a 2.7 Meuro (2.2 Meuro EU support) 6th framework Cognitive Systems STREP that will start 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 will study both fully autonomous and semi-supervised methods. The work will combine 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 will be illustrated by three demonstrators: an Image Interrogator that interactively answers simple user-defined queries about image content; a automatic annotator for people and actions in situation comedy videos; an an automatic news story summarizer. The Class consortium is interdisciplinary, combining five leading European research teams in visual recognition, text understanding & summarization, and machine learning: LEAR; Oxford University, UK; K.U. Leuven, Belgium; T.U. Darmstadt and MPI Tuebingen, Germany.

8.2.4 FP6 Network of Excellence PASCAL

Participants: Ankur Agarwal, Juliette Blanchet, Charles Bouveyron, Navneet Dalal, Gyuri Dorko, Marcin Marszalek, Jianguo Zhang, Bill Triggs, Cordelia Schmid, Frederic Jurie.

PASCAL (Pattern Analysis, Statistical Modelling and Computational Learning) is an 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 modelling 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 of LEAR coordinates the computer vision aspects and manages various activities including the Balance & Integration and Funding Review Programs. LEAR and Xerox Research Europe (XRCE) together form one of PASCAL's 14 key sites, focusing on computer vision and natural language processing.

8.2.5 FP6 Marie Curie EST Host grant VISITOR

Participants: Marcin Marszalek, Caroline Pantofaru, Cordelia Schmid, Bill Triggs, Roger Mohr.

LEAR is one of the teams participating in VISITOR, a 3 year Marie Curie Early Stage Training Host grant to the GRAVIR-IMAG laboratory to which LEAR belongs. VISITOR is funding the PhD of the Polish student Marcin Marszalek and the visit of the Canadian student Caroline Pantofaru from Carnegie Mellon University in Pittsburg.

8.2.6 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 will run from November 2005 for two years. The project aims at improving image descriptors by adding robust color information. Machine learning techniques will determine the most discriminative features, e.g. choosing between different levels of invariance and features types. Furthermore, we will learn the colorimetric properties of images and categories.

8.3 Bilateral relationships

8.3.1 University of Illinois at Urbana-Champaign, USA

Participants: Cordelia Schmid, Bill Triggs, Svetlana Lazebnik [UIUC], Jean Ponce [UIUC].

This collaboration on 3D object recognition between the research group of J. Ponce and LEAR is funded by a CNRS/INRIA/UIUC collaboration agreement. In 2005, C. Schmid visited the partner institution for one week. S. Lazebnik visited LEAR for a week and J. Ponce came twice for one day.

8.3.2 Australian National University and National ICT Australia

Participants: Bill Triggs, Richard Hartley [ANU], Alex Smola [ANU].

This collaboration centres around the Australian-funded section of the EU project LAVA, whose focuses are visual methods for recognizing particular locations and kernel based methods for visual

recognition. Bill Triggs visited ANU and NICTA (National ICT Australia) in Canberra, Australia for 3 days in August 2005 as part of this work.

9 Dissemination

9.1 Leadership within the scientific community

- Conference and workshop organization:
  - Program co-chair for IEEE Conference on Computer Vision 2005 (C. Schmid)
  - Workshops chair, some general chair duties and prize committee chair for the 10th IEEE International Conference on Computer Vision 2005 (B. Triggs)
  - Co-organizer of the MSRI Workshop on Visual Recognition, Berkeley, March 2005 (C. Schmid)
- Editorial boards:
  - International Journal of Computer Vision (C. Schmid)
  - IEEE Transactions on Pattern Analysis and Machine Intelligence (C. Schmid, B. Triggs)
  - Foundation and Trends in Computer Graphics and Vision (C. Schmid)
  - Machine Vision and Applications (R. Mohr)
- Area chairs:
  - ICCV'05 (C. Schmid, B. Triggs)
  - NIPS'05 (C. Schmid)
  - BMVC'05 (B. Triggs)
  - ECCV'06 (B. Triggs)
  - CVPR'06 (B. Triggs)
- Program committees:
  - CVPR'05 (F. Jurie, B. Triggs)
  - ICIP'05 (J. Van de Weijer)
  - ICME'05 (J. Van de Weijer)
  - NIPS'05 (B. Triggs)
  - ECCV'06 (V. Ferrari, A. Agarwal)
  - RFIA'06 (C. Schmid)
  - CVPR'06 (V. Ferrari, F. Jurie, C. Schmid)
- Other:
  - C. Schmid is a member of INRIA's Commission d'Évaluation, and of the INRIA Rhône-Alpes Comité des Emplois Scientifiques. She has participated in several recruitment committees.

  - C. Schmid is in charge of international relations at INRIA Rhône-Alpes.
  - C. Schmid was in charge of organizing the 2005 evaluation of the INRIA "CogB" research theme.
  - 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).
  - B. Triggs is a member of INRIA's COST (Scientific and Technical Strategy Committee).
  - B. Triggs is deputy director designate of the Laboratoire Jean Kuntzmann (a proposed 230 person CNRS-INRIA-INPG-UJF laboratory on applied mathematics and computer science, to be formed in 2007).
  - B. Triggs will co-direct the GDR ISIS action on machine learning and statistical methods for image understanding.
  - B. Triggs manages the Funding Review and the Balance & Integration programmes of the EU Network of Excellence PASCAL, and co-manages several other programmes.
  - Both C. Schmid and B. Triggs are among the researchers interviewed on *Computer Vision: Fact & Fiction*, a DVD produced by UC San Diego aimed at school leavers and the general public, that discusses the real state of the art in computer vision relative to that portrayed in movies.

- Prizes:

  - Ankur Agarwal received the best student paper award at the *Mini-Symposium on Machine Understanding of People and Their Responses* organized by the Rank Prize Funds (Lake District, UK, 02/2005) for his presentation "Regression models for human pose estimation from silhouettes".
  - Methods submitted by LEAR won 14 of the 18 categories of the Visual Object Recognition Challenge proposed by the European Network of Excellence PASCAL in April 2005.
  - In the ICCV'05 Computer Vision Contest, our method for matching images under wide viewpoint changes and camera localization ranked 5th among 15 and received an honorable mention. It was also by far the fastest method in the competition.

9.2 Teaching

- Matching and Recognition, INPG Masters IVR, 12 h (F. Jurie)
- Multi-media databases, INPG, 3rd year ENSIMAG, 18 h (F. Jurie)

9.3 Invited presentations

- C. Schmid. *Image description and object recognition.* University of Illinois, Urbana-Champaign, January 2005.
- A. Agarwal. *Regression models for human pose estimation from silhouettes.* Mini-Symposium on Machine Understanding of People and Their Responses, Grasmere, England, February 2005.

- B. Triggs. *Learning to reconstruct human motion from silhouettes.* Université Jean Monnet, St Etienne, France, February 2005.
- C. Schmid. *Local invariant regions and part models.* Liege University, Belgium, February 2005.
- C. Schmid. *How to define good datasets?* CVPR area chair meeting workshop, Los Angeles, February 2005.
- C. Schmid. *Modeling spatial relations.* MSRI workshop on Visual Recognition, Berkeley, March 2005.
- B. Triggs. *Regression models for human pose estimation from silhouettes.* Lund University, Sweden, April 2005.
- C. Schmid. *Recognition and matching based on local invariant features.* Cours at Oulu University, Finland, May 2005.
- C. Schmid. *Category classification.* ICCV area chair meeting workshop Leuven, Belgium, June 2005.
- B. Triggs. *Multilevel coding for visual recognition.* ICCV area chair meeting workshop Leuven, Belgium, June 2005.
- C. Schmid. *Building local part models for category-level recognition.* Empirical Inference Symposium, Tuebingen, Germany, August 2005.
- B. Triggs. *Multilevel coding for visual recognition.* Australian National University / NICTA, Canberra, Australia, August 2005.
- C. Schmid. *Building local part models for category-level recognition.* TAIMA 2005, Hammamet, Tunesie, September 2005.
- A. Agarwal. *Hyperfeatures – multilevel local coding for visual recognition.* Oxford University Robotics Research Group Seminar, Oxford, England, October 2005.
- C. Schmid. *Building local part models for category-level recognition.* ENS Ulm seminar, Paris, October 2005.
- B. Triggs. *Detecting people in images and videos and reconstructing their movements*, ENS Ulm, Paris, November 2005.
- C. Schmid. *From local invariant descriptors to recognition.* Short course at MIT, Boston, December 2005.

10 Bibliography

Books and Monographs

[1] J. PONCE, M. HEBERT, C. SCHMID, A. ZISSERMAN (editors), *Towards Category-Level Object Recognition*, Springer, To appear, `http://lear.inrialpes.fr/pubs/2006/PHSZ06`.

[2] C. SCHMID, S. SOATTO, C. TOMASI (editors), *Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, June 2005, http://lear.inrialpes.fr/pubs/2005/SST05.

Doctoral dissertations and "Habilitation" theses

[3] G. BOUCHARD, *Generative models in supervised statistical learning with applications to digital image categorization and structural reliability*, PhD thesis, Université Joseph Fourier, Grenoble, France, May 2005, http://lear.inrialpes.fr/pubs/2005/Bou05.

[4] B. TRIGGS, *Reconstruction monoculaire du mouvement humain, et autres travaux 2000-2004*, Habilitation a diriger des recherches, Institut National Polytechnique de Grenoble, Grenoble, France, January 2005, http://lear.inrialpes.fr/pubs/2005/Tri05.

Articles in referred journals and book chapters

[5] A. AGARWAL, B. TRIGGS, "Recovering 3D Human Pose from Monocular Images", *IEEE Transactions on Pattern Analysis & Machine Intelligence 28*, 1, January 2006, http://lear.inrialpes.fr/pubs/2006/AT06a.

[6] C. BOUVEYRON, S. GIRARD, C. SCHMID, "High Dimensional Discriminant Analysis", *Communications in Statistics*, Submitted, http://lear.inrialpes.fr/pubs/2005/BGS05e.

[7] C. BOUVEYRON, S. GIRARD, C. SCHMID, "High dimensional discriminant analysis", *in : Subspace, Latent Structure and Feature Selection Techniques*, Springer Lecture Notes in Computer Science, 2006, Accepted, http://lear.inrialpes.fr/pubs/2005/BGS05f.

[8] P. CARBONETTO, G. DORKÓ, C. SCHMID, H. KUCK, N. DE FREITAS, "Learning to recognize objects with little supervision", *International Journal of Computer Vision*, Submitted, http://lear.inrialpes.fr/pubs/2005/CDSKD05.

[9] M. EVERINGHAM, A. ZISSERMAN, C. K. I. WILLIAMS, L. V. GOOL, M. ALLAN, C. M. BISHOP, O. CHAPELLE, **N. Dalal**, T. DESELAERS, **G. Dorkó** , S. DUFFNER, J. EICHHORN, J. D. R. FARQUHAR, M. FRITZ, C. GARCIA, T. GRIFFITHS, **F. Jurie**, T. KEYSERS, M. KOSKELA, J. LAAKSONEN, **D. Larlus**, B. LEIBE, H. MENG, H. NEY, B. SCHIELE, **C. Schmid**, E. SEBFANN, J. SHAWE-TAYLOR, A. STORKEY, S. SZEDMAK, **B. Triggs**, I. ULUSOY, V. VIITANIBFI, **J. Zhang**, "The 2005 PASCAL Visual Object Classes Challenge", *in : Selected Proceedings of the first PASCAL Challenges Workshop*, F. d'Alche Buc, I. Dagan, and J. Quinonero (editors), LNAI, Springer, 2006, To appear, http://lear.inrialpes.fr/pubs/2006/EZKVAMCDDDDEDFGGJKKL.

[10] S. LAZEBNIK, C. SCHMID, J. PONCE, "A sparse texture representation using local affine regions", *IEEE Transactions on Pattern Analysis & Machine Intelligence 27*, 8, 2005, p. 1265–1278, http://lear.inrialpes.fr/pubs/2005/LSP05.

[11] K. MIKOLAJCZK, C. SCHMID, A. ZISSERMAN, "Human detection based on a probabilistic assembly of robust part detectors", *in : The CogViSys Project*, H.-H. Nagel (editor), Springer, 2006, To appear.

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

[13] K. MIKOLAJCZYK, T. TUYTELAARS, C. SCHMID, A. ZISSERMAN, J. MATAS, F. SCHAFFALITZKY, T. KADIR, L. V. GOOL, "A comparison of affine region detectors", *International Journal of Computer Vision 65*, 1/2, 2005, p. 43–72, http://lear.inrialpes.fr/pubs/2004/MTSZMSKG05.

[14] F. ROTHGANGER, S. LAZEBNIK, C. SCHMID, J. PONCE, "Segmenting, modeling and matching video clips containing multiple moving objects", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Submitted.

[15] F. ROTHGANGER, S. LAZEBNIK, C. SCHMID, J. PONCE, "Object modeling and recognition using local affine-invariant image descriptors and multi-view spatial contraints", *International Journal of Computer Vision 66*, 3, 2006, `http://lear.inrialpes.fr/pubs/2004/RLSP04`.

[16] C. SCHMID, G. DORKÓ, S. LAZEBNIK, K. MIKOLAJCZYK, J. PONCE, "Pattern recognition with local invariant features", *in : Handbook of Pattern Recognition and Computer Vision*, C. Chen and P. Wang (editors), edition Third, World Scientific, 2005, `http://lear.inrialpes.fr/pubs/2005/SDLMP05`.

[17] C. SMINCHISESCU, B. TRIGGS, "Building Roadmaps of Minima and Transitions in Visual Models", *International Journal of Computer Vision 61*, 1, 2005, p. 81–101, `http://lear.inrialpes.fr/pubs/2005/ST05`.

[18] C. SMINCHISESCU, B. TRIGGS, "Hyperdynamic Sampling", *Journal of Image & Vision Computing*, 2005, Special issue on ECCV'02 papers. To appear, `http://lear.inrialpes.fr/pubs/2005/ST05a`.

[19] B. TRIGGS, M. SDIKA, "Boundary Conditions for Young - van Vliet Recursive Filtering", *IEEE Transactions on Signal Processing*, 2005, To appear, `http://lear.inrialpes.fr/pubs/2005/TS05`.

[20] J. ZHANG, M. MARSZALEK, S. LAZEBNIK, C. SCHMID, "Local Features and Kernels for Classifcation of Texture and Object Categories: An In-Depth Study", *International Journal of Computer Vision*, Submitted.

Publications in Conferences and Workshops

[21] A. AGARWAL, B. TRIGGS, "Hyperfeatures – Multilevel Local Coding for Visual Recognition", Submitted to ECCV'06, `http://lear.inrialpes.fr/pubs/2005/AT05c`.

[22] A. AGARWAL, B. TRIGGS, "Monocular Human Motion Capture with a Mixture of Regressors", *in : IEEE Workshop on Vision for Human Computer Interaction at CVPR*, June 2005, `http://lear.inrialpes.fr/pubs/2005/AT05`.

[23] A. AGARWAL, B. TRIGGS, "A Local Basis Representation for Estimating Human Pose from Cluttered Images", *in : Asian Conference on Computer Vision*, January 2006. To appear., `http://lear.inrialpes.fr/pubs/2006/AT06`.

[24] J. BLANCHET, F. FORBES, C. SCHMID, "Markov Random Fields for Recognizing textures modeled by Feature Vectors", *in : International Symposium on Applied Stochastic Models and Data Analysis*, 2005, `http://lear.inrialpes.fr/pubs/2005/FBS05`.

[25] J. BLANCHET, F. FORBES, C. SCHMID, "Markov random fields for textures recognition with local invariant regions and their geometric relationships", *in : British Machine Vision Conference*, 2005, `http://lear.inrialpes.fr/pubs/2005/BFS05a`.

[26] J. BLANCHET, F. FORBES, C. SCHMID, "Modèles markoviens pour l'organisation spatiale de descripteurs d'images", *in : Conférence francophone sur l'apprentissage automatique*, 2005, `http://lear.inrialpes.fr/pubs/2005/BFS05`.

[27] J. BLANCHET, F. FORBES, C. SCHMID, "Modèles markoviens pour l'organisation spatiale de descripteurs d'images", *in : 37e Journées de Statistique de la Société Française de Statistique*, June 2005, http://lear.inrialpes.fr/pubs/2005/BFS05b.

[28] G. BOUCHARD, B. TRIGGS, "Hierarchical Part-Based Visual Object Categorization", *in : Proceedings of the Conference on Computer Vision and Pattern Recognition*, p. I 710–715, June 2005, http://lear.inrialpes.fr/pubs/2005/BT05.

[29] C. BOUVEYRON, S. GIRARD, C. SCHMID, "High Dimensional Discriminant Analysis", *in : International Symposium on Applied Stochastic Models and Data Analysis*, May 2005, http://lear.inrialpes.fr/pubs/2005/BGS05b.

[30] C. BOUVEYRON, S. GIRARD, C. SCHMID, "Une méthode de classification des données de grande dimension", *in : 37e Journées de Statistique de la Société Française de Statistique*, June 2005, http://lear.inrialpes.fr/pubs/2005/BGS05d.

[31] C. BOUVEYRON, S. GIRARD, C. SCHMID, "Une nouvelle méthode de classification pour la reconnaissance de formes", *in : 20ème colloque GRETSI sur le traitement du signal et des images*, September 2005, http://lear.inrialpes.fr/pubs/2005/BGS05c.

[32] N. DALAL, B. TRIGGS, C. SCHMID, "Human Detection using Oriented Histograms of Flow and Appearance", Submitted to ECCV'06, http://lear.inrialpes.fr/pubs/2005/DTS05.

[33] N. DALAL, B. TRIGGS, "Histograms of Oriented Gradients for Human Detection", *in : Proceedings of the Conference on Computer Vision and Pattern Recognition, 2*, p. 886–893, June 2005, http://lear.inrialpes.fr/pubs/2005/DT05.

[34] G. DORKO, C. SCHMID, "Maximally Stable Local Description for Scale Selection", Submitted to ECCV'06.

[35] J. FALCOU, J. SEROT, T. CHATEAU, F. JURIE, "A Parallel Implementation of a 3D Reconstruction Algorithm for Real-Time Vision", *in : Parallel Computing*, 2005, http://lear.inrialpes.fr/pubs/2005/FSCJ05.

[36] F. JURIE, B. TRIGGS, "Creating Efficient Codebooks for Visual Recognition", *in : International Conference on Computer Vision*, 2005, http://lear.inrialpes.fr/pubs/2005/JT05.

[37] D. LARLUS, G. DORKÓ, F. JURIE, "Création de Vocabulaires Visuels Efficaces pour la Catégorisation d'Images", *in : Reconnaissance des Formes et Intelligence Artificielle*, 2006, http://lear.inrialpes.fr/pubs/2006/LDJ06.

[38] S. LAZEBNIK, C. SCHMID, J. PONCE, "A Maximum Entropy Framework for Part-Based Texture and Object Recognition", *in : International Conference on Computer Vision*, 2005, http://lear.inrialpes.fr/pubs/2005/LSP05a.

[39] L. MASSON, M. DHOME, F. JURIE, "Tracking 3D Object using Flexible Models", *in : British Machine Vision Conference*, 2005, http://lear.inrialpes.fr/pubs/2005/MDJ05.

[40] T. MOERLAND, F. JURIE, "Learned Color Constancy From Local Correspondences", *in : IEEE International Conference on Multimedia & Expo*, 2005, http://lear.inrialpes.fr/pubs/2005/MJ05.

[41] E. NOWAK, F. JURIE, "Vehicle Categorization: Parts for Speed and Accuracy", *in : VS-PETS workshop at ICCV*, 2005, http://lear.inrialpes.fr/pubs/2005/NJ05.

[42] J. PONCE, T. PAPADOPOULO, M. TEILLAUD, B. TRIGGS, "On the Absolute Quadric Complex and its Application to Autocalibration", *in : Proceedings of the Conference on Computer Vision and Pattern Recognition*, p. I 780–787, June 2005, http://lear.inrialpes.fr/pubs/2005/PPTT05.

[43] J. VANDEWEIJER, C. SCHMID, "Coloring local feature extraction", Submitted to ECCV'06.

Internal Reports

[44] A. AGARWAL, B. TRIGGS, "Hyperfeatures - Multilevel Local Coding for Visual Recognition", *Research Report number RR-5655*, INRIA Rhone-Alpes, August 2005, http://lear.inrialpes.fr/pubs/2005/AT05b.

[45] C. BOUVEYRON, S. GIRARD, C. SCHMID, "Analyse Discriminante de Haute Dimension", *Research Report number RR-5470*, INRIA Rhone-Alpes, January 2005, http://lear.inrialpes.fr/pubs/2005/BGS05.

[46] G. DORKÓ, C. SCHMID, "Object Class Recognition Using Discriminative Local Features", *Research Report number RR-5497*, INRIA Rhone-Alpes, February 2005, http://lear.inrialpes.fr/pubs/2005/DS05a.

[47] J. ZHANG, M. MARSZALEK, S. LAZEBNIK, C. SCHMID, "Local Features and Kernels for Classification of Texture and Object Categories: An In-Depth Study", *Research report*, INRIA Rhone-Alpes, October 2005.

Miscellaneous

[48] N. DALAL, "HOG Object Detection - version 0.0.1 du 12/07/2005", July 2005, Copyright owned by INRIA. Software registered at Agence pour la Protection des Programmes (APP) 249, rue de Crimée-75019 Paris, France., http://lear.inrialpes.fr/pubs/2005/Dal05.