Project/Team LEAR : Learning and Recognition in Vision

Frédéric Jurie, Cordelia Schmid, Bill Triggs

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team LEAR

Learning and Recognition in Vision

Rhône-Alpes

THEME COG

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]

Vice-head and scientific co-director

Bill Triggs [CR1, CNRS]

Research member

Frédéric Jurie [CR1, CNRS]

Faculty member

Roger Mohr [Professor, ENSIMAG]

Administrative assistant

Anne Pasteur

Post-doctoral fellow

Jianguo Zhang [INRIA scholarship, 12/2003-03/2005]

Technical staff

Michaël Sdika [09/2003-03/2005]

PhD students

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

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

Guillaume Bouchard [UJF, INRIA scholarship until 11/2004]

Charles Bouveyron [UJF, MENESR scholarship, co-supervised with INRIA project MISTIS]

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

Navneet Dalal [INPG, INRIA scholarship]

Gyuri Dorkó [INPG, INRIA scholarship]

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

Student interns

Ankur Agarwal [DEA IVR INPG, INRIA scholarship, 11/2002–10/2004]

Juliette Blanchet [DEA de Mathématiques, Université de Toulouse, 03/2004–08/2004]

Aurélie Bugeau [Masters IVR INPG, 03/2004–07/2004]

Peter Carbonetto [Masters U. Vancouver, INRIA scholarship, 10/2003–06/2004]

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

Diane Larlus-Larrondo [TER UFRIMA, 02/2004–06/2004, then Masters, 10/2004–08/2005]

Tijmen Moerland [Masters, 02/2004–10/2004]

2 Overall Objectives

LEAR's main focus of research is learning based approaches to visual object recognition and scene interpretation, particularly for image retrieval and video indexing. Understanding the content of everyday images and videos is one of the most challenging problems in computer vision. The extent to which we can do this is currently limited, but we believe that very significant advances will be made over the next few years by combining emerging statistical learning techniques with state of the art image descriptors. This field is also close to a major threshold of applicability: even partial solutions are likely to enable many new applications.

LEAR's main research areas are:

- **Image description.** Many efficient lighting and viewpoint invariant image descriptors are now available, such as for example affine-invariant interest points. Our current research aims to extend these techniques to describe textures, to define more powerful similarity and saliency measures and to characterize 2D and 3D shape information.

- **Learning.** Our research on machine learning and statistical modeling is mainly aimed at improving their applicability to visual recognition and computer vision. It 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: dealing with the huge amounts of data that image and video collections contain; handling large rich natural class hierarchies rather than just simple yes/no classifiers; and capturing enough information about the domain to allow generalization from just a few images, rather than from large carefully marked-up training databases.

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

3 Scientific Foundations

3.1 Image description

We believe that the extraction of robust image descriptors is a critical component of any visual recognition system, and even though many efficient descriptors are already available, further research is clearly needed in this area. One can go a certain distance using simplistic descriptors, but their unreliability and lack of invariance puts a heavy burden on the learning method and the training data and ultimately limits the performance that can be achieved. Better descriptors allow simpler learning methods to be used and produce better separation of classes, potentially allowing generalization from just a few examples instead of requiring large, carefully engineered training databases.

The kinds of descriptors that we advocate have a certain number of basic properties:

- **Locality and redundancy:** For resistance to changes of background and occlusions, reduced sensitivity to changes of viewpoint and variable intra-class geometry, and robustness against individual feature extraction failures, descriptors should have relatively small spatial support, but 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.

- **Salience:** Fragments are not very useful unless they can be extracted automatically and found again in other images. Hence, rather than using general fragments, we focus on local descriptors based at particularly salient points — “keypoints” or “points of interest”. This gives a sparser and hence 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.

- **Photometric and geometric invariance:** The interest points and their descriptors should have an appropriate degree of invariance to changes of illumination and variations of local image geometry induced by changes of viewpoint, viewing distance, and local intra-class variability. In practice, geometric invariance is usually approximated by invariance to Euclidean, similarity or affine transforms of the local image.

- **Informativeness:** Notwithstanding all of the above types of invariance, the descriptors should 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 this requires relatively high dimensionality. Just as importantly, the useful information should be manifest, not hidden in obscure high-order correlations between coefficients. Image formation is essentially a spatial process, so in practice this favours descriptors that code relative position information manifestly (e.g. context-style descriptors rather than moments or Fourier descriptors).

Our current research in this area is focused on creating detectors and descriptors that are better adapted to particular kinds of imagery, incorporating spatial neighborhood and region constraints to improve informativeness, and extending the scheme to cover different kinds of locality.

3.2 Learning

We are interested in learning and statistics mainly as technologies for attacking difficult vision problems, so we take an eclectic approach, using a wide variety 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 controlled by various types of regularization, model and feature selection, and dimensionality reduction methods, after being measured using methods such as cross-validation, information criteria and capacity bounds. Visual descriptors tend to be high dimensional and they typically contain some redundancy, so we often preprocess data using techniques such as PCA and its nonlinear variants, ICA, and LLE/Isomap, to reduce it to a more manageable dimensionality. 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. Missing data is common owing to unknown class labels, feature detection failures, occlusions and intra-class variability, so we often need to use completion techniques such as Expectation Maximization. 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 contain huge amounts of data, so we need to use algorithms suited to large-scale learning problems, and it is expensive and tedious to label large numbers of training images, so unsupervised, semi-supervised and transductive learning methods are of particular interest. 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, correlational, and co-learning are useful for handling this.

We keep up to date on learning technology by maintaining active links with both the statistics community, most notably via collaborations with the INRIA projects MISTIS and SELECT (formerly IS2), and the machine learning one, most notably via the EU project LAVA and the Network of Excellence PASCAL.

3.3 Recognition

The current state of progress in visual recognition shows clearly that combining advanced image descriptors with modern learning and statistical modelling techniques has the potential to produce very significant advances. We believe that, taken together and tightly integrated, these techniques have the potential to make visual recognition a widespread technology.

The kind of process that we advocate makes full use of the unusual robustness and richness of our image description methods (see §3.1) to provide a vocabulary of base features that already goes a long way towards characterizing the category being recognized. The final learning based classifier is thus mainly responsible for extending the model to larger amounts of intra-class variation and gross changes of aspect or viewpoint, and for capturing the subtler higher-order correlations that are needed to fine tune the base performance. That said, our approach is not simply feature extraction *then* learning: the integration is actually much tighter than this. Nearly every stage of our descriptor chain uses learning and statistical modelling in a fundamental way, to generate or select robust invariant features, to squeeze out redundancy and bring out informativeness. Similarly, to maximize their performance, the final learning methods use descriptor comparison metrics (kernels, reference densities, structural models) that are intimately based on the statistical properties and invariances (or lack thereof) of the learned descriptors.

4 Application Domains

A solution to the general object recognition problem will enable a wide range of applications including defense, health care, human-computer interaction, image retrieval and data mining, industrial and personal robotics, manufacturing, scientific image analysis, space exploration, surveillance and security, and transportation. In fact, with the ever expanding array of image sources, some form of automatic object recognition technology must eventually be an integral part of every information system. Even partial solutions are likely to enable many applications.

One of LEAR's main application domains is image and video indexing. This is an area with huge potential. For example, it is estimated that 96% of all data currently generated by humanity is personal images and home videos[1]. Currently, we are working on developing indexing techniques for camera equipped hand-held devices such as personal digital assistants, on object-level structuring and indexing of feature film videos, and on applying our techniques to surveillance in the context of military applications.

Object-level image and video structuring organize the content of an image or a video in terms of the objects and actions contained in it, and thus allow the user to browse and access it in terms of high-level, semantically meaningful units. For example, within the European project VIBES, software was developed that allows a video database to be searched for a given set of characters, scenes or actions

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

based on user-selected regions containing examples of these. In the European project AceMedia, we are developing methods that find humans in still images and categorize their actions.

A personal visual assistant is a portable device equipped with a camera that can identify the category or instance of an object that it sees, and supply the user with associated information. A software prototype is being developed within the European project LAVA to test and validate our algorithms.

Surveillance requires the detection and recognition of objects. In our case, a military application, the camera is static and the detection is therefore relatively straightforward. The subsequent recognition should then differentiate between different types of vehicles.

5 Software

5.1 Software for computing local invariant features

Participants: Cordelia Schmid, Michael Sdika, Gyuri Dorko, Krystian Mikolajczyk [U. Oxford], Andrew Zisserman [U. Oxford], Tinne Tuytelaars [KU Leuven], Luc Van Gool [KU Leuven], Jiri Matas [U. Prague].

The local feature extraction programs developed during K. Mikolajczyk's PhD at INRIA and his postdoc at Oxford [3, 4, 22] have been improved in speed and accuracy. A well-documented library is available to the members of the team. A joint evaluation has also been made, to compare the performance of the approach with affine invariant feature extractors developed at Oxford, Leuven and Prague [30]. The study is based on a carefully designed test setup with well-defined comparison criteria and a set of images containing representative scenes viewed under different transformations. The software is available on the internet and should allow the evaluation of future detectors. The executables of the different detectors included in the comparison are also available, on `http://lear.inrialpes.fr/sotfware`.

5.2 Object recognition demonstrator

Participants: Cordelia Schmid, Michael Sdika, Bill Triggs, Roger Mohr.

By combining affine invariant feature extraction with a descriptor retrieval mechanism and a verification step, an image database can be searched for examples of a given specific object or scene. The demonstrator runs under Linux with a web camera. A retrieval search takes about half a second in a database containing 400 images. Figure 1 shows an example of the interface.

5.3 Human detection software

Participants: Navneet Dalal, Bill Triggs, Cordelia Schmid.

Software for detecting humans in static images has been developed as part of the European Union FP6 Integrated Project AceMedia. The current version of the method is restricted to images of fully visible, standing people. A new database containing more than 1800 annotated human images has also been collected to allow the method to be tested.

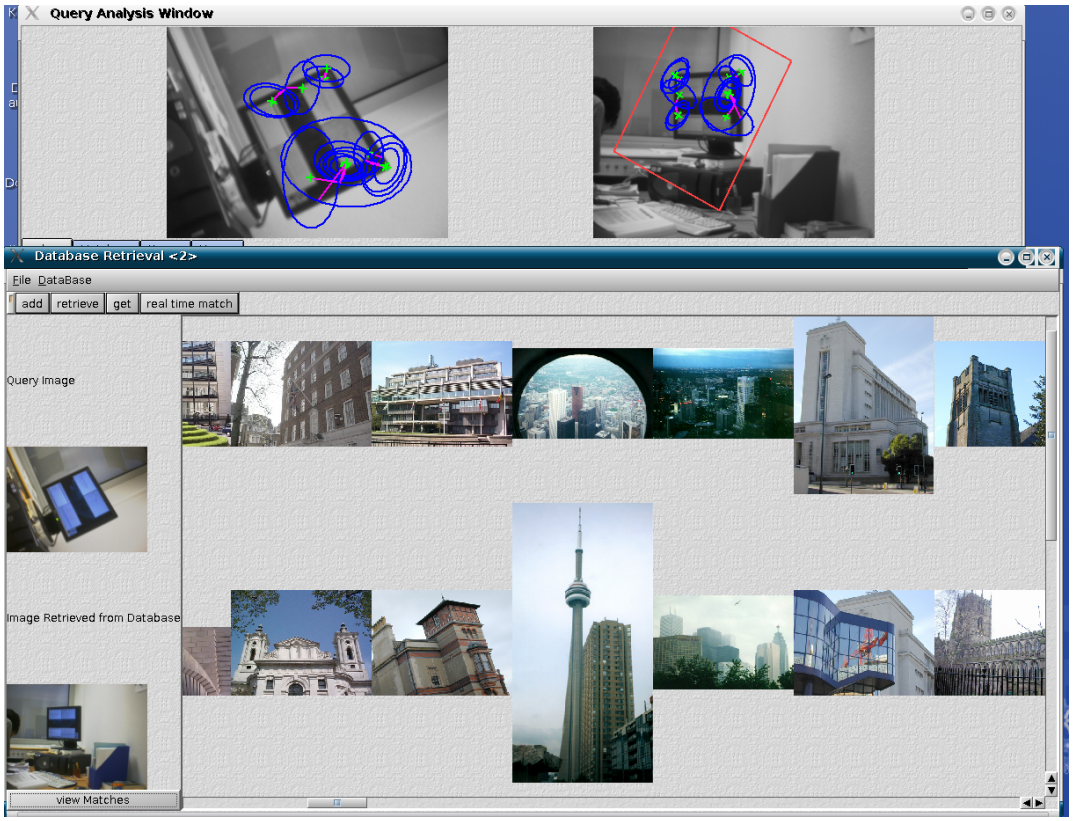


Figure 1: Top row: the matched affine-invariant regions and the estimated transformation between the query and the retrieved image. Note that the regions recover the same photometric information despite large changes of viewpoint, and hence allow rich viewpoint-invariant descriptors to be calculated. Bottom row: on the left the query image and the retrieved image. On the right some of the other database images.

6 New Results

6.1 Image description

Participants: Cordelia Schmid, Frédéric Jurie, Bill Triggs, Michael Sdika, Krystian Mikolajczyk [Oxford], Andrew Zisserman [Oxford], Salil Jain, Aurélie Bugeau, Tijmen Moerland.

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

6.1.1 Affine-invariant descriptors

We have developed scale- and affine-invariant salient point detectors [1, 4] that give excellent performance for recognizing both specific objects and scenes, and texture and object classes [6, 7].

Scale invariance is obtained by searching for maxima in scale-space. Different functions can be used to construct the scale-space, for example the Laplacian or the Hessian of the image intensity. A combination of Harris interest points computed in scale-space with a scale selection based on the Laplacian has shown very good performance. However, in the presence of significant viewpoint changes, scale invariance alone does not suffice for reliable recognition, and an extension to affine invariance is necessary. This is obtained by running an iterative affine warping procedure that reduces the interest point's second-moment matrix to normal form. For each point we then obtain an associated affine-invariant region on which a conventional descriptor can be computed (see figure 1 above). 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, the Leuven detectors combining points and edges as well as extracting intensity maxima, the Prague one extracting maximally stable connected components. We have collaborated with Leuven, Oxford and Prague on a comparison of these approaches. The results [30] 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. For example some are more adapted to structured scenes and others to textures. In most of our experiments, either the Prague MSER regions or our Hessian-Affine points provide the best repeatability score, followed by the Harris-Affine points. In contrast, the Oxford salient regions give relatively low repeatability. For the Leuven edge-based region detector, the performance largely depends on scene content, i.e. whether the image contains stable curves or not. The Leuven intensity extrema based region detector gives average scores. Another contribution of our study is the carefully designed test setup (c.f. section 5).

6.1.2 Performance evaluation of local descriptors

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 have evaluated the pairing of several different descriptors with several different interest point detectors and have developed an improved descriptor [3], see also[MS03] for preliminary results.

Our evaluation was carried out for different 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.

We compared shape context, steerable filters, PCA-SIFT, differential invariants, spin images, SIFT, complex filters, moment invariants, and cross-correlation for different types of interest regions. We

[MS03] K. MIKOLAJCZYK, C. SCHMID, "A performance evaluation of local descriptors", *in : IEEE Conf. Computer Vision & Pattern Recognition, 2*, p. 257–263, June 2003.

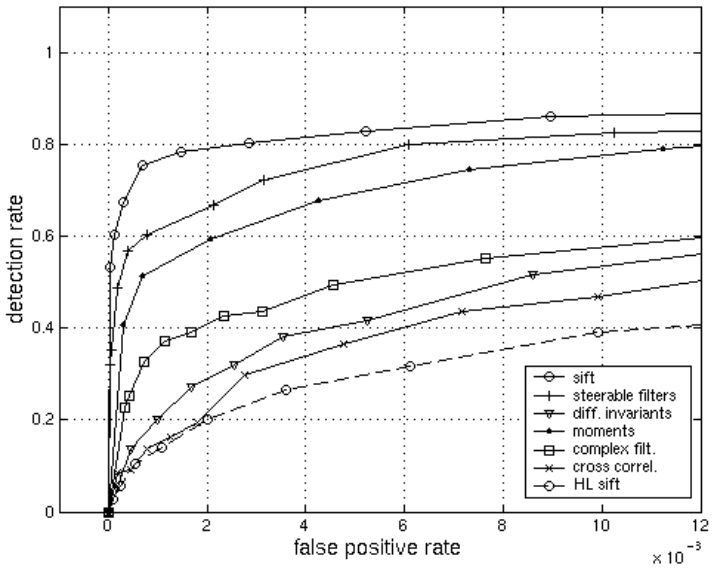


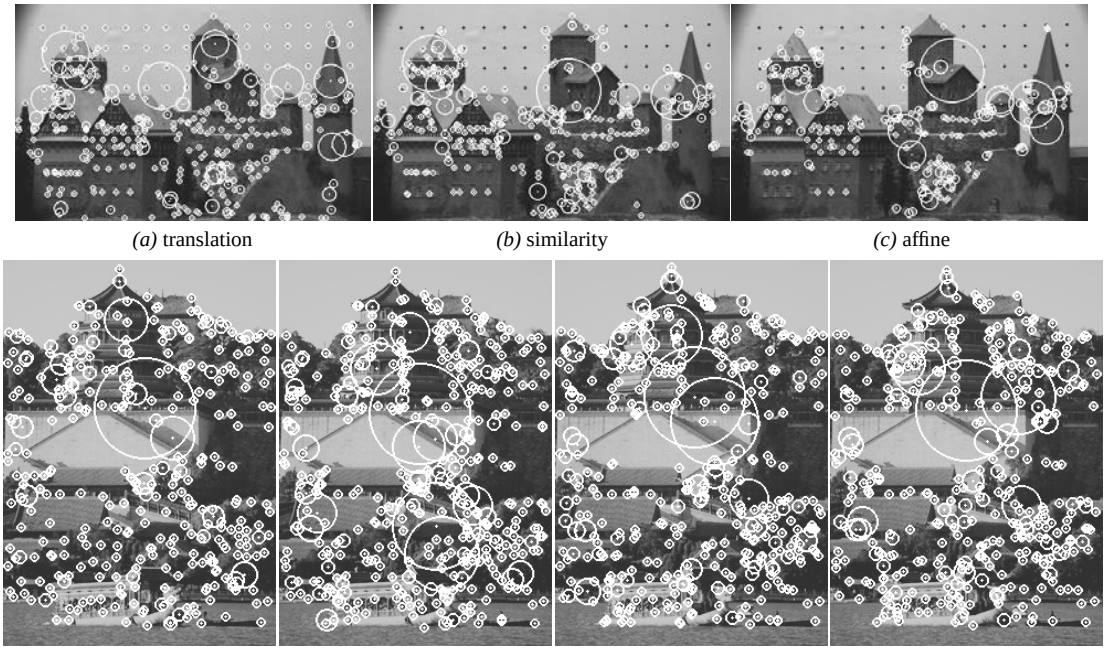
Figure 2: Descriptor evaluation for a camera viewpoint change of $60°$, using affine-invariant Harris regions. `HL sift` is the SIFT descriptor computed for scale-invariant Harris regions.

also proposed an extension of the SIFT descriptor, and showed that it outperforms the original method. The ranking of the descriptors is mostly independent of the point detector used, with the SIFT based descriptors performing best. Their success can be explained by their robustness against localization errors and small geometric distortions. Moments and steerable filters show the best performance among the low dimensional descriptors.

Some typical results for a significant viewpoint change between the query and database image are shown in figure 2, using the affine-invariant Harris-Laplace detector. The SIFT descriptors are clearly the most robust, but note that the differences between descriptors are less significant than the improvement produced by ensuring affine invariance: SIFT descriptors computed with only *scale*-invariant Harris-Laplace regions ('`HL SIFT`' in the figure) perform worse than any of the affine descriptors shown here, as the underlying regions are not invariant enough to be re-detected reliably. Steerable filters come second, but they perform significantly worse than SIFT descriptors here.

6.1.3 Indexing of visual descriptors

In order to perform image retrieval or visual object recognition based on SIFT and similar high-dimensional 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 high-dimensional neighbouring point or region search – a problem that is currently for practical purposes essentially linear in the database size, and hence slow for large databases. We have developed an approximate method to speed up such 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 k-D 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.



(a) translation *(b)* similarity *(c)* affine

(d) translation *(e)* translation + rotation *(f)* translation + scale *(g)* similarity

Figure 3: Some examples of keypoints found by our detectors. The detectors have selected points that have maximal local stability under the indicated transformations. For example, the dots in the background of the top image provide good stability against translations but poor stability against rotations, so they are selected as keypoints under the translation group, but not under the group of similarity transformations.

6.1.4 Detecting keypoints with stable position, orientation and scale under illumination changes

Local feature approaches to vision geometry and object recognition are based on selecting and matching sparse sets of visually salient image points ('keypoints' or 'points of interest'). Their performance depends critically on the reliability with which corresponding keypoints can be found in subsequent images. Among the many existing keypoint selection criteria, the popular Förstner-Harris approach explicitly targets geometric stability, defining keypoints to be points that have locally maximal self-matching precision under translational least squares template matching. However, many applications require stability in orientation and scale as well as in position. Current approaches detect translational keypoints and try to impose good orientation and scale behaviour afterwards. This is suboptimal, and can be misleading when different motion variables interact. We have developed a more principled formulation, based on extending the Förstner-Harris approach to general motion models [25]. The method also incorporates a simple local appearance model to ensure good resistance to the most common illumination variations. Some examples of detections with the new approach under various different motion models are shown in figure 3.

Figure 4: Some examples of salient regions output by our multiscale attention model.

6.1.5 Multiscale model of visual attention

Visual salience — the selection of especially informative or "interesting" image regions as focuses for further computation — is central to our philosophy of visual recognition. In the past we have mainly used salience measures based on local feature detectors developed in computer vision. In the work described here, we take another approach to salience — the human vision motivated models, among which the work by Itti & Koch is perhaps the best known example — and generalise it to a multiscale model that is potentially suitable for use in our multiscale object recognition framework. The method is based on calculating image pyramids encoding local contrast measures for intensity, colour and texture and combining them heuristically into an overall saliency map. The original approach produced a single saliency image, whereas here we modify the saliency metric and evaluate it at multiple scales to produce a saliency pyramid that encodes the scale as well as the position of the most salient image regions. Figure 4 shows several examples of the salient regions detected. The work is reported in the Masters thesis [28].

6.1.6 Shape features

The image features presented in the first two sections are based on grey-level image information. Local invariant features based on such information have proven to be very successful for matching and recognition of specific textured objects. Unfortunately, for many objects the only reliable recognition cues are edges or shape, and texture cannot be used as the primary descriptor. In particular, for category-level recognition, edge and shape are often the only reliable common features between different instances of the category. To cover this case, we have recently developed two different types of scale-invariant edge descriptors.

The first, developed in 2003, characterizes the type of edge pixels by the edge information in their neighbourhoods [MZS03]. Each edge region is an edge pixel around which the distribution of relative positions and orientations of its surrounding edges are described.

In contrast, the *local shape features* do not centre the region on an edge pixel, but instead extract features in regions of influence that capture the local structure (shape) of the surrounding contour image. Our approach [18] detects local shape convexities by local scale-space maximization of a robust concentricity measure, the entropy of radial gradient orientations in an annulus whose radius

[MZS03] K. MIKOLAJCZYK, A. ZISSERMAN, C. SCHMID, "Shape recognition with edge-based features", *in: British Machine Vision Conference, 2*, p. 779–788, September 2003.

% Precision	97.4	93.1	90.2	87.7	77.3	60.5
% Recall	23.2	46.7	63.4	70.4	82.5	96.4

Figure 5: Horse detection results. Top: detections in a few test images. Bottom : precision and recall for several detection thresholds.

defines the scale. This is robust to clutter inside the annulus, occlusions, and spurious edge detections. It has been combined with a simple geometric model and a voting process to allow the detection of objects. Figure 5 shows some results for horse recognition and detection and some quantitative results in terms of precision (the percentage of detections that are correct) and recall (the percentage of correct images that are retrieved).

6.1.7 Color description

Color is a useful local image descriptor, but it is sensitive to illumination changes and so cannot be used without an initial normalisation process (color constancy), which is difficult to provide in the general case where the lighting is unknown. Our approach 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.

Our approach uses the *color flow model* to describe lighting changes between images depicting the same scene. 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 color changes between pairs of images, see figure 6.

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 normalisation 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,

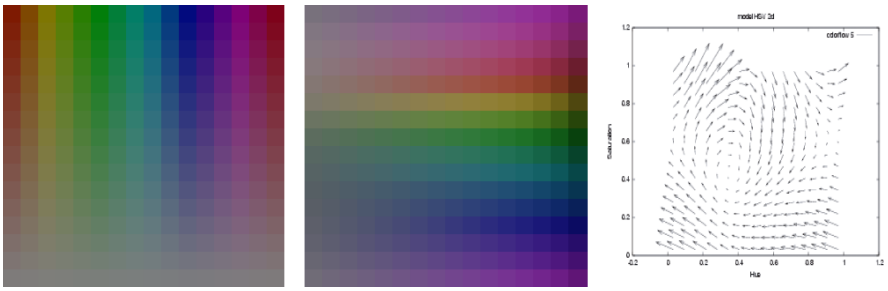


Figure 6: An illustration of the color flow model. Left : the reference color map (Hue / Saturation representation). Middle : the color map for a new illumination of the scene. Left : the transformation between the two maps (vectors represent individual color changes).

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.

6.2 Learning

Participants: Bill Triggs, Cordelia Schmid, Guillaume Bouchard, Charles Bouveyron, Peter Carbonetto, Juliette Blanchet.

Keywords: Discriminative-generative learning, Gaussian mixture models, Support Vector Machines, semi-supervised learning, data reduction.

6.2.1 Discriminative versus Generative Learning

In statistics there are two main approaches to classification. *Discriminative* methods try to learn a single model that directly predicts the class label from the observed input data, whereas *generative* ones learn separate models for each class, then choose the model (class) that best fits the observed data. Generative models are more general, they extend more naturally to complex problems (missing data, multiple classes…), and they are often stabler and simpler to learn because the classes do not interact. However discriminative models typically have better absolute classification performance as they are optimized directly for this, and they can avoid modelling details that are important for class description but irrelevant for inter-class discrimination.

We have developed a method that captures some of the advantages of each approach by combining them. Technically, this is done by fitting a discriminative model (i.e. optimizing classification performance over all classes together) that includes a generative-model-based penalty term to enforce some of the regularization (but also some of the mismodelling) implicit in the full generative model. The strength of the penalty is set by cross-validation. The combined model often has higher performance than either of its parents. The work was published in the IASC International Symposium on Computational Statistics, COMPSTAT 2004 [16].

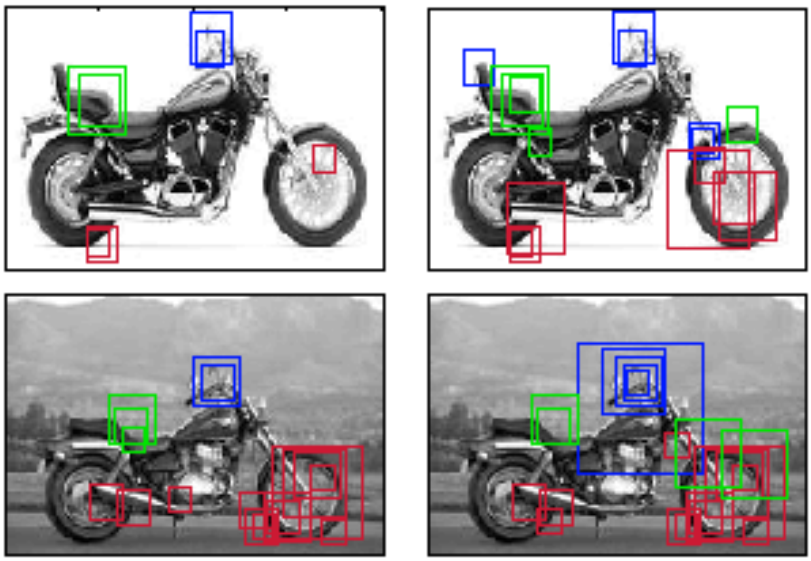


Figure 7: Recognition of motorbike parts using HDDA (left) and SVM (right). The colors blue, red and green are respectively associated with the handlebar, wheel and seat features.

6.2.2 Dimensionality reduction

Visual features and descriptors are often high dimensional, which leads to overfitting and thus penalizes classification and recognition accuracy [17]. Global dimensionality reduction is not always useful because although the descriptors for a given visual class often concentrate on a low dimensional subspace of descriptor space, the descriptors of different classes belong to quite different subspaces (often with different dimensionalities). We have introduced a new method, called High Dimensional Discriminant Analysis (HHDA), that reduces the dimension of each class independently and regularizes the class conditional covariance matrices in order to adapt the Gaussian framework to high dimensional data. The regularization is achieved by assuming that each class has spherical Gaussian deviations from its eigenspace. It is possible to make additional assumptions to further reduce the number of parameters that are estimated.

Experimental results comparing our approach to classical dimensionality reduction techniques show that HDDA performs significantly better. A further comparison with SVM (support vector machine – designed to handle high dimensions) also demonstrates the performance of HDDA. Figure 7 presents an example of part classification for motorbikes. The results on the left were obtained with HDDA and those on the right with SVM. On the top row, for the example, HDDA recognizes motorbike parts without any errors while SVM makes five misclassifications.

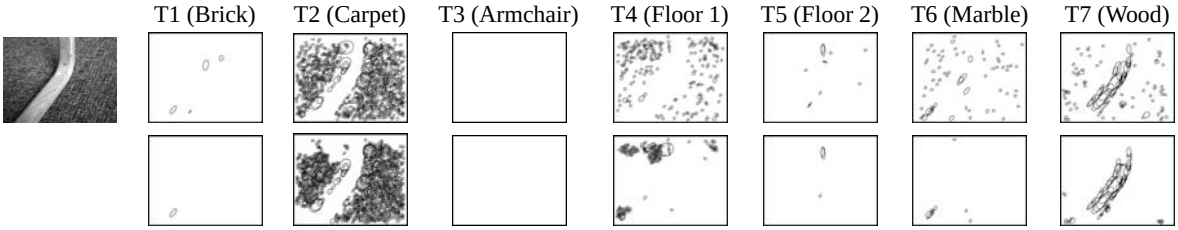


Figure 8: Segmentation/classification results for natural texture recognition. Image regions are labeled using maximum likelihood (top) and mean field (bottom). Adding spatial constraints improves the results.

6.2.3 Markov random fields for recognizing textures

Our previous work on texture recognition [LSP03a] 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.

The co-occurrence statistics do not explicitly model the dependencies between neighboring descriptors – they assume that they are statistically independent variables. Recognition results can be further improved by modeling the dependence between descriptors [27]. Here we use a parametric dependency model based on Markov Random Fields (MRF's). These models require non trivial parameter estimation: we use recent methods based on the mean field principle of statistical physics. Using sample images, textures are learned as MRF's and a set of estimated parameters is associated with each texture. At recognition time, another MRF is used to compute, for each feature vector, the membership probabilities for the different texture classes. Preliminary experiments on seven natural texture classes (each with 20 training samples) show promising results, see figure 8. The regions in the different columns correspond to the classification results. Note that adding spatial constraints reduces the number of misclassified regions.

6.2.4 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 spent 8 months in LEAR applying UBC's approach for constrained semi-supervised learning by data association to the selection of discriminative local features. The setup is the same as in section 6.3.2: 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 generalised Gibbs sampler

[LSP03a] 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.

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 initialisation. 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 factorially with the number of components.

The experimental results for classification of test images on the dataset described in section 6.3.2 are on average about 2 percent better than the ones obtained there [29]. However learning the full data association is orders of magnitude more time consuming, and the results are also influenced by the values of the constraints.

6.3 Recognition

Participants: Cordelia Schmid, Bill Triggs, Frederic Jurie, Roger Mohr, Ankur Agarwal, Guillaume Bouchard, Gyuri Dorko, Navneet Dalal, Salil Jain, Jianguo Zhang, Svetlana Lazebnik [UIUC], Fred Rothganger [UIUC], Jean Ponce [UIUC], Krystian Mikolajczyk [Oxford], Andrew Zisserman [Oxford], Cristian Sminchisescu [Toronto].

Keywords: visual models, object class recognition and detection, human detection.

6.3.1 Texture recognition

In our past work, we developed a method for weakly supervised learning of visual models that can handle both complex natural textures — for example "textured" animals such as zebras and leopards — and highly structured patterns such as parts of a face [8]. The visual model is based on a two-layer image description: a set of underlying "generic" descriptors, and a learned distribution over neighbourhoods. This description method is rotationally invariant, robust to model deformations, and it efficiently characterizes "appearance-based" visual structure. The learning method is based on selecting distinctive clusters in descriptor space — descriptors common in the positive and rare in the negative examples — and using these as features for object recognition. Experimental results for "textured" animals and faces show very good performance for both retrieval and localization.

This representation has now been extended to allow recognition of texture patterns despite appearance variations due to non-rigid transformations and changes in viewpoint [2], see also[LSP03b] for preliminary results.

At the feature extraction stage, a set of affine-invariant local patches is extracted from the image. Affine-invariant patches provide both a good level of spatial selectivity, and a texture representation that is invariant to any geometric transformation that can be locally approximated by the affine model. Each patch is characterized using a gray-level descriptor. In each image, the distribution of descriptors is summarized as a set of weighted clusters. These signatures are then compared using the Earth Mover's Distance (EMD), a convenient and effective dissimilarity measure. For our application, the signature/EMD framework offers several important advantages. Signatures are more descriptive than

[LSP03b] S. LAZEBNIK, C. SCHMID, J. PONCE, "Sparse Texture Representation Using Affine-Invariant Neighborhoods", *in : IEEE Conf. Computer Vision & Pattern Recognition, 2*, p. 319–324, 2003.

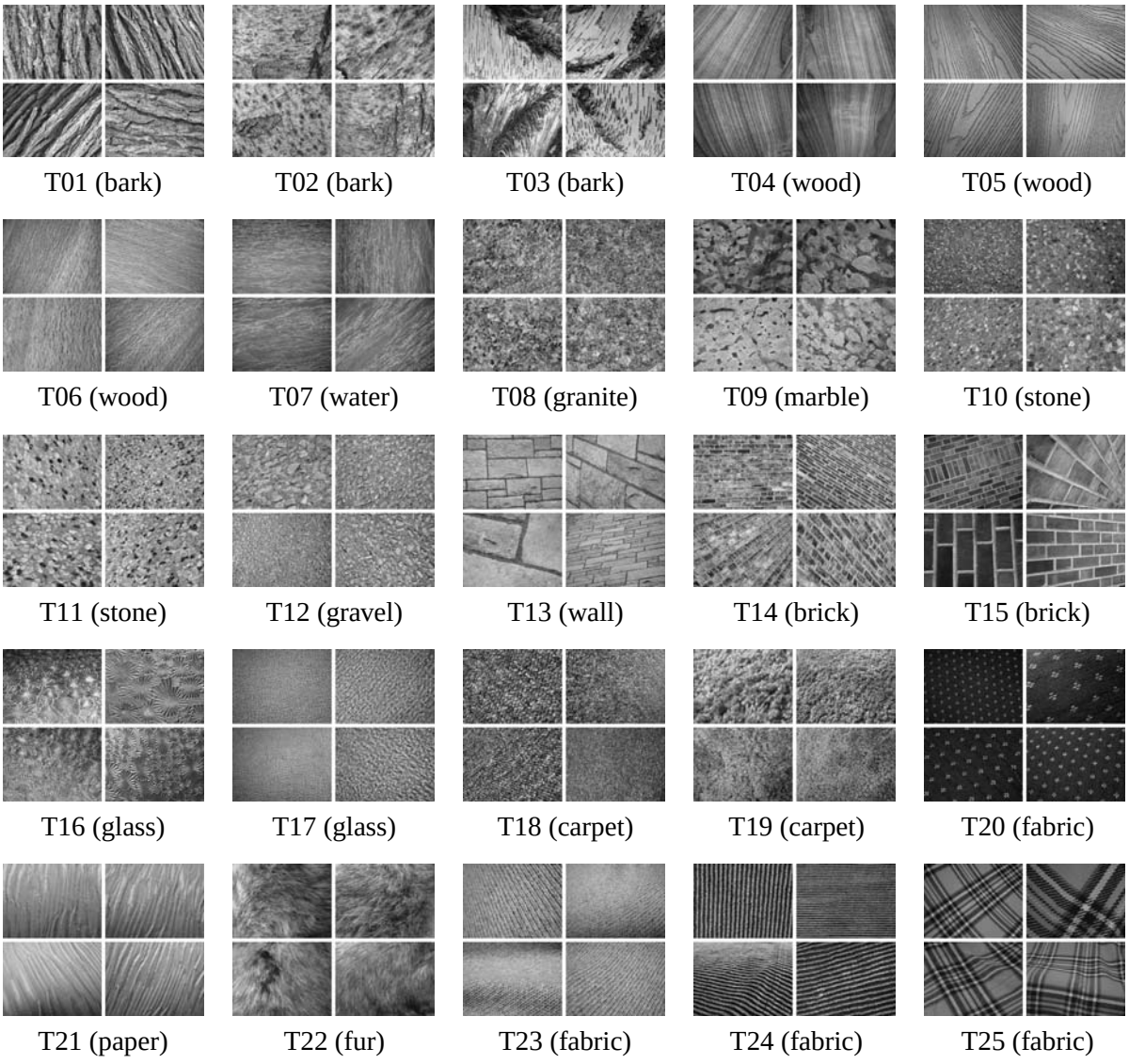


Figure 9: Four samples each of the 25 texture classes used in the experiments.

histograms with fixed bins and do not require global clustering of the descriptors from all images. EMD can match signatures of different sizes, and it is not very sensitive to the number of clusters. This texture representation has been evaluated for retrieval and classification tasks using a collection of images of textured surfaces taken from different viewpoints. Figure 9 shows four sample images from each texture class. Significant viewpoint and scale changes occur within each class, and several of the classes include additional sources of variability: inhomogeneities in the texture patterns, non-rigid transformations, illumination changes, and unmodeled viewpoint-dependent appearance changes. Retrieval and classification results for these textures are excellent [2].

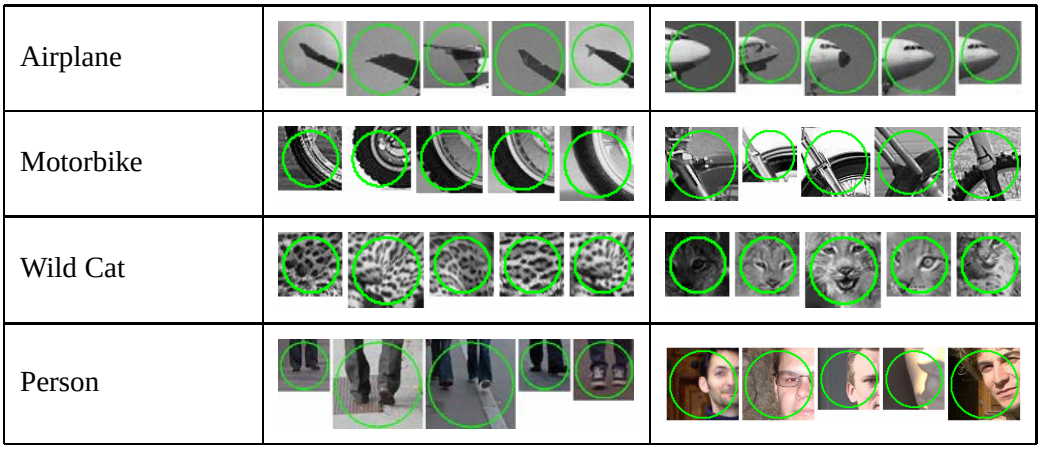


Figure 10: Examples of discriminative parts (clusters) for the categories airplane, motorbike, wild cat and person. Parts are chosen from the 10 most discriminative ones. Each cluster is represented by five of its members. The circles represent the regions extracted by the scale-invariant detector.

6.3.2 Recognition of object classes – feature selection

Learning and recognizing object class models from images is one of the most difficult problems in computer vision. The combination of image description and machine learning/pattern classification techniques has recently led to significant progress. Several recent approaches to category-level object recognition use classification techniques to acquire part models from training images, and then train a classifier to recognize objects. This paradigm has been successfully applied to the recognition of cars, faces and human beings in complex imagery. However, the image descriptors used in these methods enjoy very limited invariance properties (mostly translational invariance), which severely limits the range of admissible viewpoints that they can handle. This can be avoided by using local invariants as image descriptors.

Our approach finds groups (clusters) of similar scale-invariant local features, i.e. local parts, and selects among these parts the ones which best discriminate between positive and negative images [29], see also [DS03] for preliminary results. The experimental results demonstrate the power of using invariant local features for building discriminative local parts. Figure 10 shows two discriminative parts for the categories airplane, motorbike, wild cat and person.

The approach is weakly supervised, that is the training images are labeled as positive and negative, but the objects are not labeled in the positive images. The training set is split into a clustering set and a validation set. Parts are learned based on the clustering set and the significance of each part is then determined with the validation set.

The first step of our approach is to extract local invariant features. Here we use Harris-Laplace and Harris-Affine as well as the entropy detector by Kadir and Brady and describe the regions by the SIFT descriptor. The descriptors from the clustering set are then used to learn the individual parts. We estimate the Gaussian mixture model of their distribution and each component of the mixture represents a part (cluster). The EM algorithm is used to estimate the parameters of the mixture model and

[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.

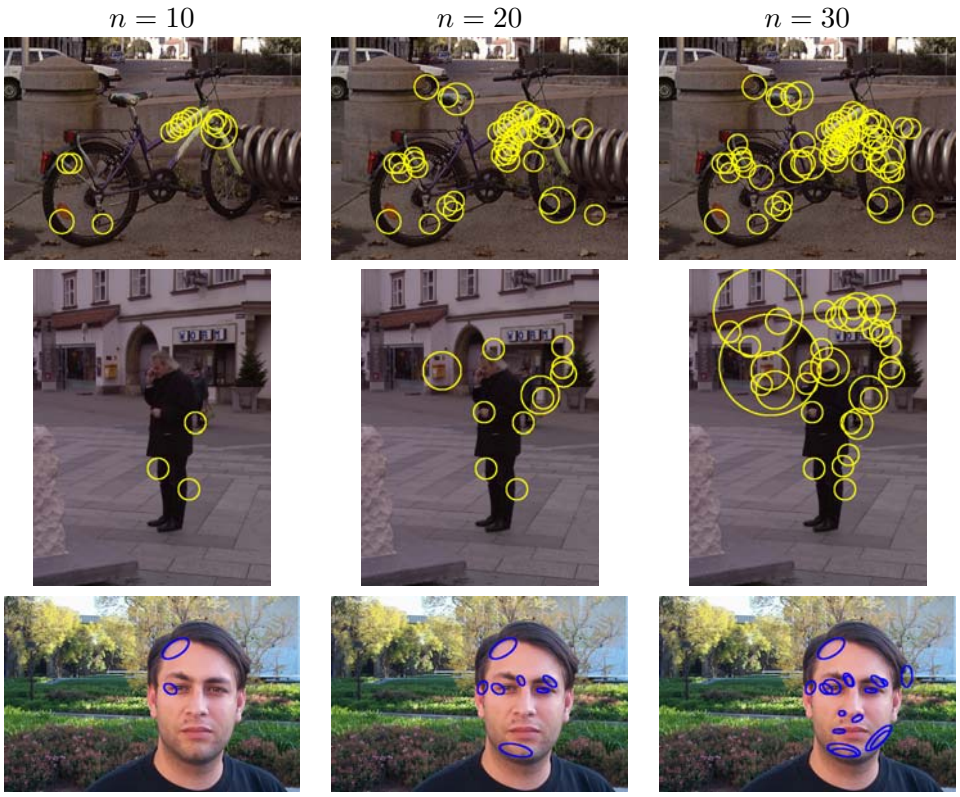


Figure 11: Positive detections with increasing n for different categories and detectors. First and second row: entropy detector. Third row: Harris-Affine detector.

is initialized with the output of the K-means algorithm. Descriptors are assigned to the components (parts) with the maximum posteriori probability. We then select parts with the validation set. Each part is ranked by the likelihood ratio between the descriptors of the positive images – note that the individual descriptors are unlabeled – and the negative images. Other criteria can be used, such as mutual information [DS03]. The final classifier then sets the n highest ranked parts as positive and the others as negative. A descriptor is classified as an object descriptor if it is labeled as belonging to one of the top n parts (maximum posteriori probability for that class) where n is a parameter of our approach.

Our approach is evaluated in two different ways. We first verify that the positive descriptors lie mostly on the object. Figure 11 shows the results for a few test images. Only a few points are incorrectly classified, and they could easily be eliminated by applying a simple spatial coherence criterion.

We then evaluate the performance by image classification, that is whether or not the image contains the object. This is a standard criterion that allows comparison with existing methods. We compare image classification results in table 1. The training and test images are the same as in Fergus, Perona & Zisserman (2003) and Opelt, Fussenegger, Pinz & Auer (2004). We measure performance with the Receiver Operating Characteristic (ROC) equal error rate, *i.e.* the point on the ROC curve – obtained

[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.

	Our model	Fergus *et al*	Opelt *et al*
airplanes	0.985	0.902	0.889
faces	0.991	0.964	0.935
motorbikes	0.995	0.925	0.922
wildcats	0.87	0.900	—
bikes	0.88	—	0.865
people	0.88	—	0.808

Table 1: Image classification performance measured at equal error rates.

by varying the number of parts n – where the proportion of true positives is equal to the proportion of true negatives. Classification requires an additional parameter, namely the minimum number p of positive descriptors in the image for which the image is classified as positive. This is estimated from the validation set. Table 1 shows that our model using the two scale-invariant detectors Harris-Laplace and the entropy detector outperforms the other methods.

6.3.3 Recognition of object classes – part based models

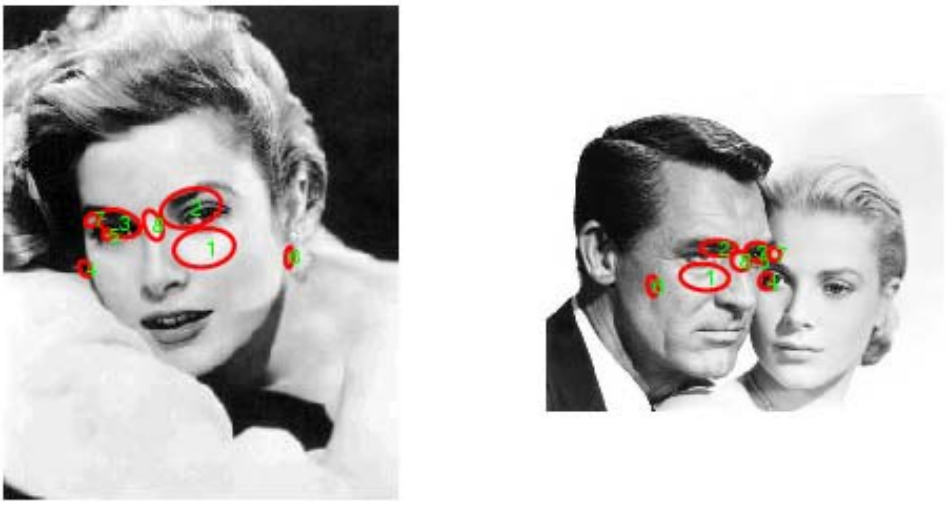


Figure 12: Matching faces: an affine-invariant semi-local part.

In this section we use characteristic patterns formed from affine-invariant patches linked by semi-local spatial relations to describe salient object parts [19, 20]. Figure 12 illustrates this idea with an affine-invariant semi-local part found between face images using the output of the affine-invariant Laplacian and a variant of affine alignment. Note that the part is stable despite large viewpoint variations and appearance changes.

Affine-invariant semi-local parts are geometrically stable configurations of multiple affine-invariant regions, found by the affine Laplace detector. These parts are approximately affinely rigid by construction, i.e. the mapping between instances of the same part in two images can be well represented by

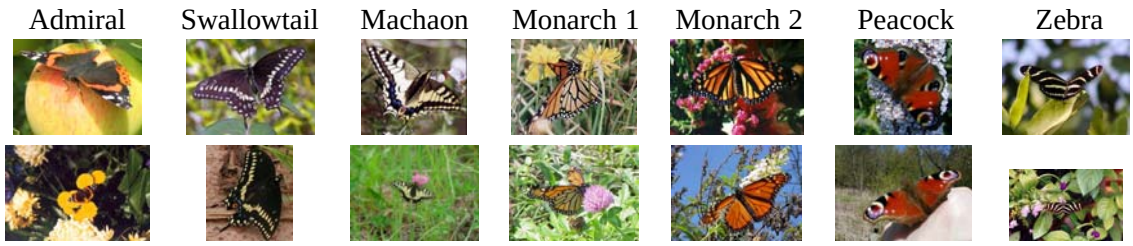


Figure 13: The butterfly dataset. Two samples of each class are shown in each column.

a 2D affine transformation. Combined with the *locality* of the parts, this property makes our method suitable for modeling a wide range of 3D transformations, including viewpoint changes and non-rigid deformations. Furthermore, they are more distinctive than the individual features used in the previous section. The mechanism for learning affine-invariant semi-local parts is based on the idea that direct search for visual correspondence is the key to successful recognition. Thus, at training time we seek to identify groups of neighboring affine regions whose appearance and spatial configuration remains stable across multiple instances. To avoid the prohibitive complexity of establishing simultaneous correspondence across the whole training set, we separate the problem into two stages: parts are first initialized by matching pairs of images, then matched against a larger *validation set*. Even though finding optimal correspondences between features in two images is still intractable, effective sub-optimal solutions can be found using non-exhaustive constrained search.

We have conducted experiments for the challenging application of automatic acquisition and recognition of butterfly models in heavily cluttered natural images. Figure 13 shows two samples for the seven classes in our dataset of 619 butterfly images. No negative images are used. The pictures were collected from the Internet. They are extremely diverse in terms of size and quality. Motion blur, lack of focus, re-sampling and compression artifacts are common. The dataset is particularly appropriate for illustrating the descriptive power of semi-local affine parts, because the geometry of a butterfly is locally planar for each wing, but not globally planar. In addition, the species identity of a butterfly is determined by a basically stable geometric wing pattern, though appearance can be significantly affected by variations between individuals, lighting, and imaging conditions. This problem is beyond the capabilities of many state-of-the-art recognition systems.

The recognition framework is straightforward. Matching and validation are used to identify a fixed-size collection of parts for representing the classes. Candidate parts are formed by matching between eight randomly chosen pairs of training images. Ten verification images per class are used to rank candidate parts according to their repeatability score, and the top ten parts per class are retained for recognition. Figure 14(a) shows the part with the highest repeatability for each of the classes. At testing time, the parts for all classes are detected in each training image. If multiple instances of the same part are found, we retain only the single instance with highest number of detected regions. Figure 14 (b) shows examples of part detections in individual test images. The score for a class is defined as the cumulative *relative repeatability* of all its parts, or the total number of regions detected in all parts of the classes divided by the sum of part sizes. For multi-class classification, each image is assigned to the class having the maximum score. Figure 15(a) shows classification results obtained using the above approach (the average rate is 90.4%), and Figure 15(b) shows how performance is improved by using multiple parts.

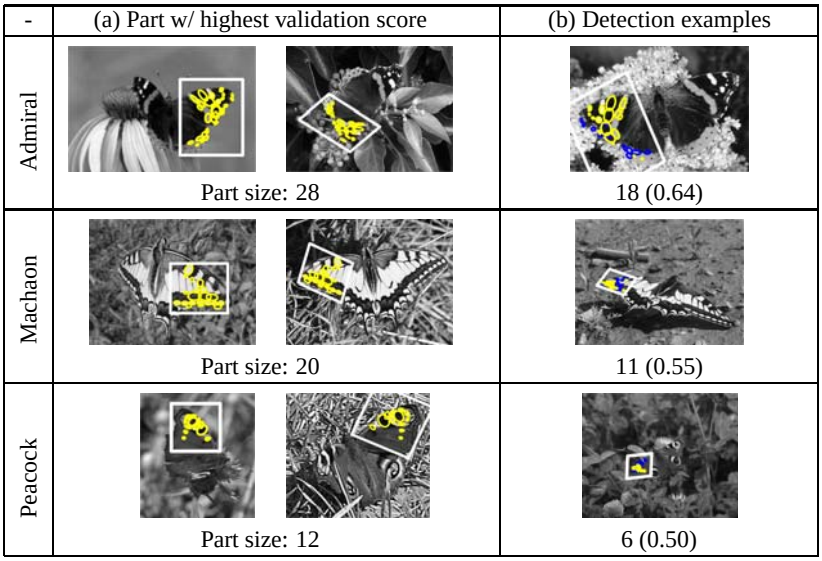


Figure 14: Butterfly modeling and detection examples. (a) The part with the highest validation score for a class. The part size is listed below each modeling pair. (b) Example of detecting the part from (a) in a single test image. Detected regions are shown in yellow and occluded ones are reprojected from the model in blue. The total number of detected regions (absolute repeatability) and the corresponding repeatability ratio are shown below each image.

(a)

Class	Part size		Test images	Correct	(rate)
Admiral	179	(12/28)	85	74	(0.871)
Swallowtail	252	(18/29)	16	12	(0.750)
Machaon	148	(12/21)	57	55	(0.965)
Monarch 1	289	(14/67)	48	35	(0.729)
Monarch 2	275	(19/36)	58	53	(0.914)
Peacock	102	(8/14)	108	108	(1.000)
Zebra	209	(16/31)	65	58	(0.892)
Total			437	395	(0.904)

(b)

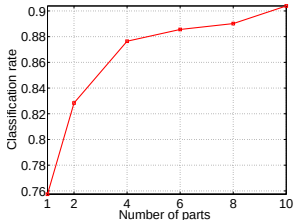


Figure 15: Classification results for the butterflies. (a) The second column shows the size of the model (top 10 parts) for each class (the sum of sizes of individual parts), and the size of the smallest and the largest parts are listed in parentheses. (b) Classification rate versus number of parts.

6.3.4 Recognition of object classes – hierarchical models

The parts-based models described above are based on combining features in relatively rigid arrangements. A different class of approaches is “bag of features” methods, which simply select some characteristic classes of local image features and count how many of them occur in a given image region. Despite their lack of spatial structure, these approaches prove quite effective for many object categories, especially ones with characteristic texture and weak spatial structure. We have developed an approach intermediate between the two, which combines local feature classes with a loose spatial model that is capable of capturing the variations present in many generic object categories.

The method is based on a generative probabilistic model that codes the geometry and appearance of visual object categories as a loose hierarchy of parts, with probabilistic spatial relations linking parts

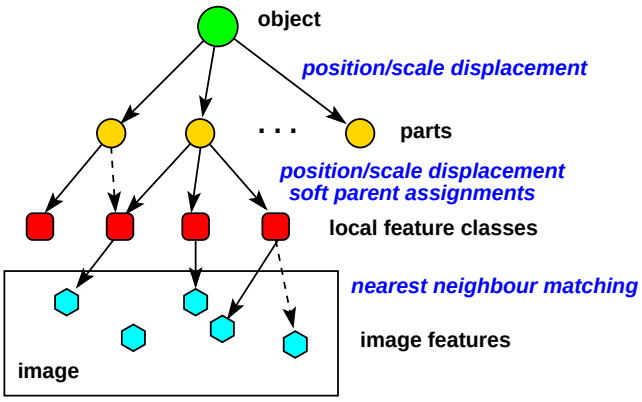


Figure 16: An illustration of the structure of our hierarchical part based model for visual object recognition. The model is a tree-structured hierarchy of parts and subparts with the complete object at the root and individual scale-invariant local feature classes at the leaves. A probability distribution 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.

to subparts, soft assignment of subparts to parts, and scale invariant keypoint based local features at the lowest level of the hierarchy. It efficiently handles models containing hundreds of redundant local feature classes, such as those returned by current keypoint detectors. The high 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 16 sketches the structure of the model, and figure 17 shows examples of how it adapts to changes of viewpoints and the intra-class variations of a generic visual object class.

6.3.5 Building 3D models from multi-view description

It is well-established that a set of images of a rigid 3D object can be used for object recognition. However, purely image based representations are not optimal as a great deal of redundant information must be stored and they do not provide a 3D model that can be used for verification. We build a 3D model from the images and use it for recognition [5]. We use the affine-invariant patches introduced in section 6.1.1 to represent local surface appearance, and select promising matches between pairs of images or an object model and an image. We then use the geometric multi-view consistency constraints studied in the structure-from-motion literature to represent the global object structure, retain correct matches, and discard incorrect ones. Our experiments show that rigid object models can be acquired automatically from a few images (figure 18), and then used effectively for recognition tasks (figure 19).

In order to obtain a quantitative comparison of our method with other state-of-the-art object recog-

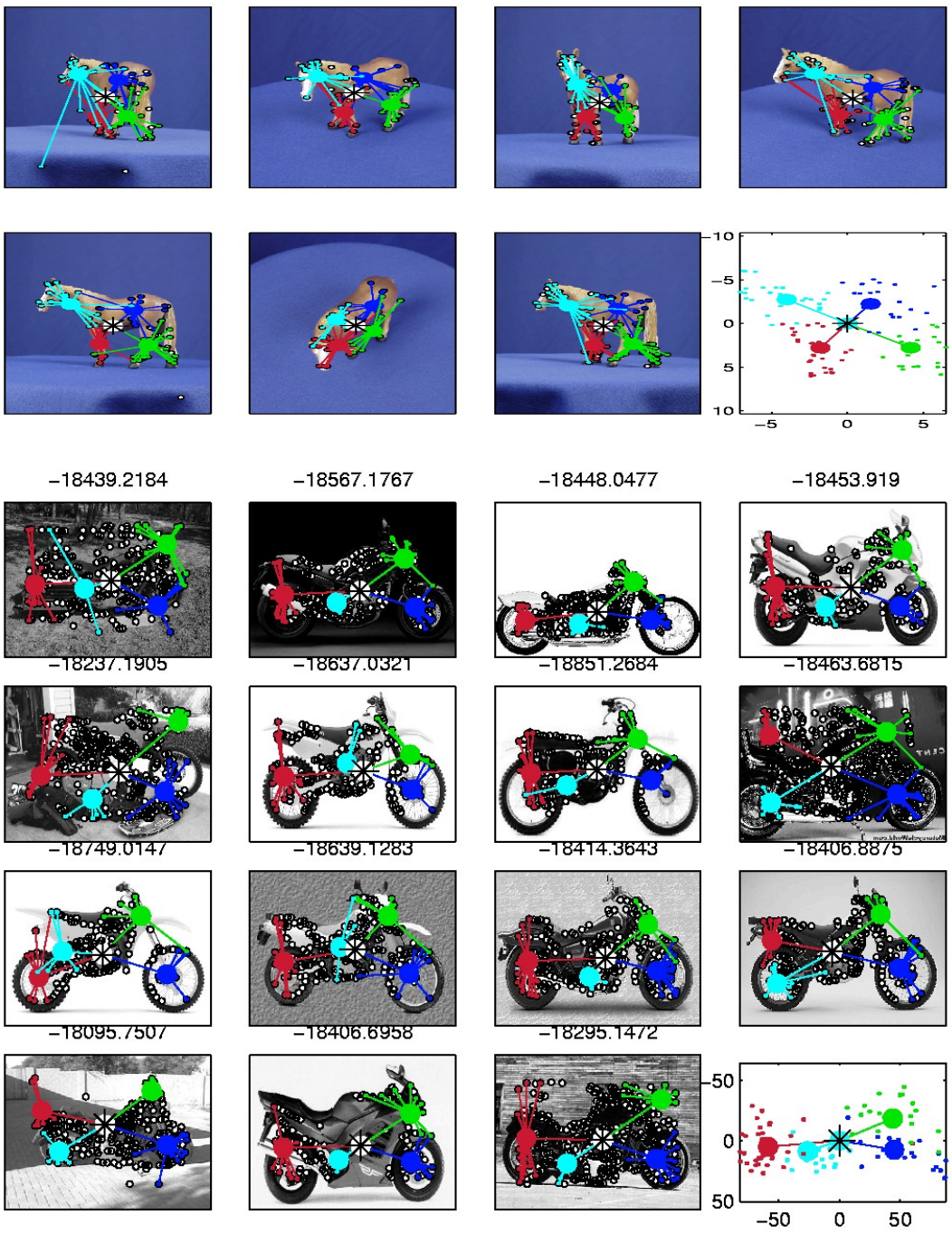


Figure 17: Some examples of our hierarchical model for object recognition in action. The first set of images shows that owing to its loose hierarchical structure, the model has good resistance to viewpoint changes. The second set shows that it can adapt to class variations, here different types of motorcycle.

Figure 18: Object gallery: sample input images and renderings of the corresponding models.

Figure 19: Results of a recognition experiment. Left: A test image. Right: Instances of five models (a teddy bear, a doll stand, a salt can, a toy truck and a vase) have been recognized, and the models are rendered in the poses estimated by our program. Bounding boxes for the reprojections are shown as black rectangles.

nition systems, we have provided our dataset to several other research groups. The algorithms proposed by Ferrari, Tuytelaars & Van Gool (2004), Lowe (2004), Mahamud & Hebert (2003), and Moreels, Maire & Perona (2004) have been tested by their authors in this comparative study. As shown by Figure 20, all the algorithms perform well on our data set, achieving recognition rates of 90% and above for false detection rates below 10%. In this experiment, the color version of our algorithm and Lowe's program perform best for very low false detection rates, followed by the black-and-white version of our algorithm. The technique proposed by Ferrari et al. achieves an extremely high recognition rate at the cost of a somewhat higher false detection rate. Although all five algorithms use multiple views to form object models, only Lowe's algorithm and ours actually combine the information associated with multiple views in the recognition process. Note that Lowe's algorithm does not construct an ex-

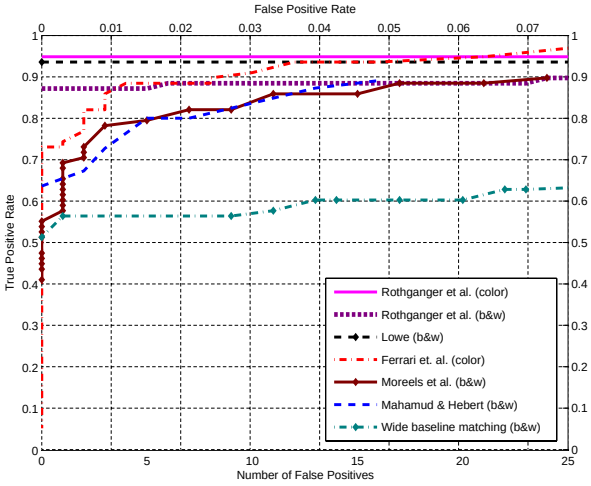


Figure 20: True positive rate plotted against number of false positives for several different recognition methods, see text for details.

plicit 3D model, but it allows multiple training views sharing common patches to vote for the same object. The other methods consider all training pictures independently, which essentially reduces object recognition to image matching. These results provide evidence that combining multiple views improves recognition performance, as does the inclusion of geometric constraints in the matching process. Of course, there is a price to pay for the integration of multiple images into a single model. First, it makes modeling more costly and complicated. Second, it requires the use of training views with sufficient overlap.

The 3D object models can also be extracted from video sequences [24], i.e. for dynamic scenes composed of multiple independently moving rigid objects. Multi-view constraints associated with groups of affine-covariant scene patches, together with a normalized description of their appearance, are used to segment a scene into its rigid parts, construct three-dimensional models of these parts, and match instances of models recovered from different image sequences. This allows moving objects to be detected and recognized in video sequences. This has been applied to the task of *shot matching*, i.e., the identification of shots that depict the same scene in a video clip.

6.3.6 Human detection – image descriptors

As part of our ongoing work on human detection, we have made a detailed comparative study of different image descriptors for this problem. “Pedestrian detection” (the detection of fully visible standing or walking people) was selected as the standard test problem, and a monolithic (non-parts-based) linear Support Vector Machine run in a moving rectangular image window was used as standardized detection framework giving rapid detection and good baseline performance. The families of image descriptors tested included Haar and Harr-like Gaussian wavelets, shape contexts, and various Histogram of Oriented image Gradient (HOG) based descriptors, inspired by the success of SIFT descriptors [Low04]

[Low04] D. LOWE, “Distinctive Image Features from Scale-invariant Keypoints”, *International Journal of Computer Vision 60*, 2, 2004, p. 91–110.

Figure 21: Some sample images from our new human detection database.

but here used in a dense uniform grid instead of being sampled only sparsely at salient local feature points. The HOG based feature sets proved to be best, giving false positive rates around two orders of magnitude lower than the best existing wavelet based detectors. Optimization over the various descriptor parameters showed 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 of the system. As part of this work, we also developed a new test database containing more than 1800 annotated positive images – see figure 21. Figure 22 summarizes the performance of the different descriptors, and figure 23 shows which features within the detector window are most important for the detection process.

6.3.7 Human detection – combination of classifiers

Our previous work on human detection [RST02] was based on detecting individual body parts and assembling them using dynamic programming. It showed good results in simple conditions, but lacked robustness in general settings. Our new approach [21] models not individual body members, but rather larger regions that include some local context. This increases distinctiveness, while still remaining local enough to allow for occlusion and the detection of close-up views. This is not the case for state-of-art pedestrian detectors based on global information. We also use robust part detectors based on gradient and Laplacian local features that efficiently capture the shape information. Using the probabilistic co-occurrence of these features increases their distinctiveness without reducing robustness. Learning with AdaBoost combines features with the highest co-occurrence probabilities. The detection results are further improved by computing a probabilistic score that takes the relative positions of the parts in the assembly into account. The approach is also very efficient as (i) all part detectors

[RST02] R. RONFARD, C. SCHMID, B. TRIGGS, "Learning to Parse Pictures of People", *in: European Conference on Computer Vision*, p. IV 700–714, Copenhagen, 2002, `http://lear.inrialpes.fr/pubs/2002/RST02`.

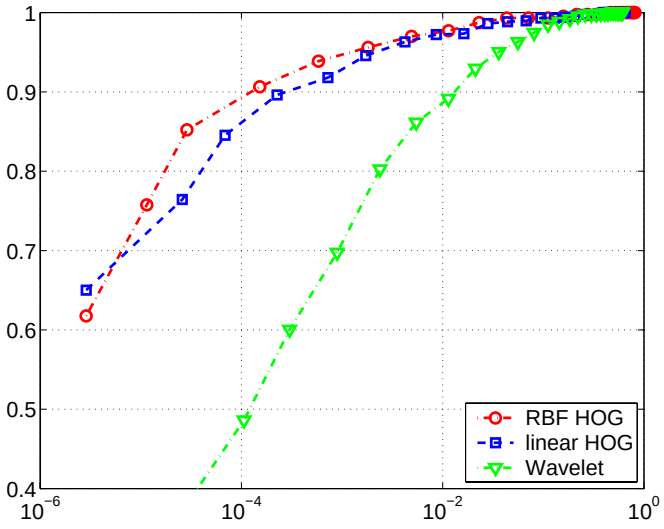


Figure 22: A summary of the performance of different types of descriptor sets on our new test database. Histogram of Oriented Gradient (HOG) features have false positive rates around two orders of magnitude lower than the best wavelet features. Using a Gaussian kernel SVM in place of a linear SVM further increases the descriptor performance (by about 3% at 10^{-4} false positives per window tested), but at the cost of much higher running time.

use the same initial features, (ii) a coarse-to-fine cascade approach is used for part detection, (iii) an assembly strategy reduces the number of spurious detections and the search space. The results (see figure 24) are very promising and outperform existing human detectors. Furthermore, the face detection results for frontal and profile views, obtained as one part of the human detector, are comparable with state-of-art detectors.

6.3.8 Human Tracking and Action Recognition

Over the past few years we investigated several approaches to the problem of human motion capture (the estimation of articulated human pose and movement) from monocular images and image sequences. This year we have published motion capture work based on: explicit 3-D body models and generative image modelling [9, 10]; dynamical modelling for 2-D image based body tracking [14]; and learning based 3-D motion capture from image silhouettes, without explicit body modelling [11, 13, 26].

Model based 3-D motion capture: First consider approaches based on explicit 3-D body models and generative image modelling. Our previous work in this area established the fact that local minima of the parameter-space likelihood induced by the kinematic structure of the problem are a major barrier to reliable tracking, and we have developed several non-convex optimization techniques designed to counter this problem. This year we have continued this work with journal versions of two approaches that find nearby minima in the cost (likelihood) function by first finding the “mountain passes” (codi-

Figure 23: An illustration of the image information encoded by the HOG based detectors. In each triplet, we display from left to right: (1) the input image; (2) the corresponding HOG feature vector (only the dominant orientation of each cell is shown); (3) the orientations that dominate the descriptor output (obtained by multiplying the feature vector with the weights learned for the linear SVM). Inspection shows that for an appropriate choice of descriptor parameters, the SVM is able to map positive example images to a set of canonical edges corresponding to human contours.

mension one saddle points) that lead to them. The first uses techniques based on modified forms of local Newton minimization to find the saddle points [10], the second, *hyperdynamic sampling*, uses a modified Markov Chain Monte Carlo sampler that focuses samples near saddles [9]. Figure 25 illustrates the results of a search for local minima produced by one of our Newton based techniques.

Learning dynamical models for 2D articulated human body tracking: Another strategy is to avoid 3D ambiguities by tracking only *2D* articulated image motion in the first instance. However a 2D formulation can not express 3D rigidity and viewpoint invariance constraints, and this leads to its own set of ambiguities. To strengthen the model enough to permit reliable tracking, it is useful to include priors characterizing "typical" 2D human pose and dynamics. In practice these have quite complex forms so they must be learned from training images. We have continued our previous year's work on this subject with a paper [14] focused on dynamical modelling. The method learns a piecewise linear autoregressive model by self-consistent clustering of a set of marked-up training sequences. Body parameters (34D vectors of 2D joint angles and body lengths) are taken from the training data, reduced to 5–8 dimensions by linear PCA to stabilize later estimation steps, and partitioned using K-means. A linear autoregressive dynamical model is learned for each cluster and lifted to give 34-D predictions. Then the data is reclustered according to regularized model prediction accuracy, the models are re-learned, and the whole process is repeated to convergence in an EM-like loop. The result is a flexible piecewise model that is capable of capturing the details of human dynamics, even

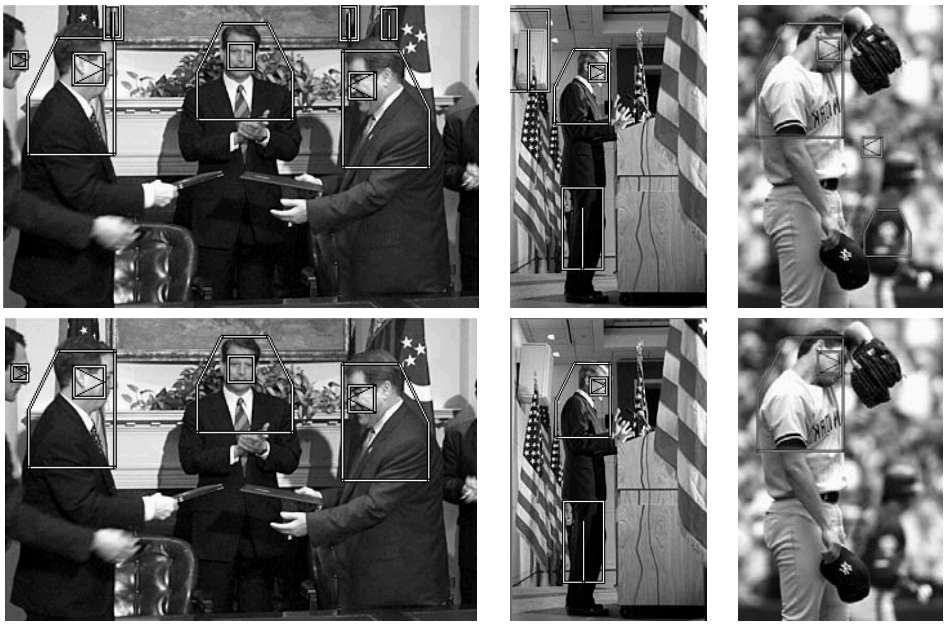


Figure 24: Results for human detection. Top row: body part detection. Bottom row: detection with the joint likelihood model. Note that the joint likelihood model significantly improves the detection results.

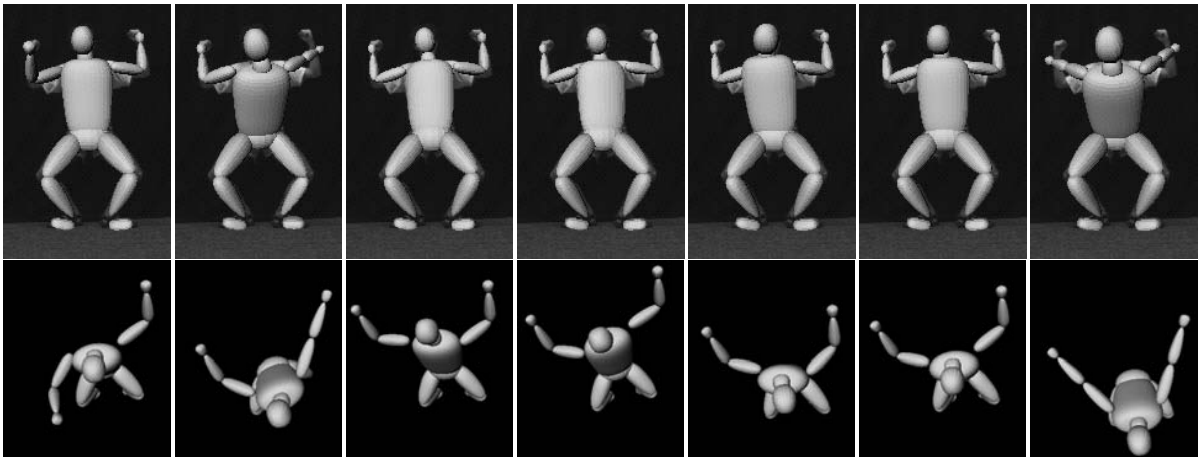


Figure 25: Some examples of 'reflective' local minima induced by the kinematic structure of the problem of motion capture from monocular images. The minima were found by *eigenvector tracking*, one of our Newton based search methods for nearby local minimum. The images show the human model from the original camera viewpoint and from a synthetic overhead viewpoint. Note the pronounced forwards-backwards character of these reflective minima, and the fact that in each case the model fits the original image well.

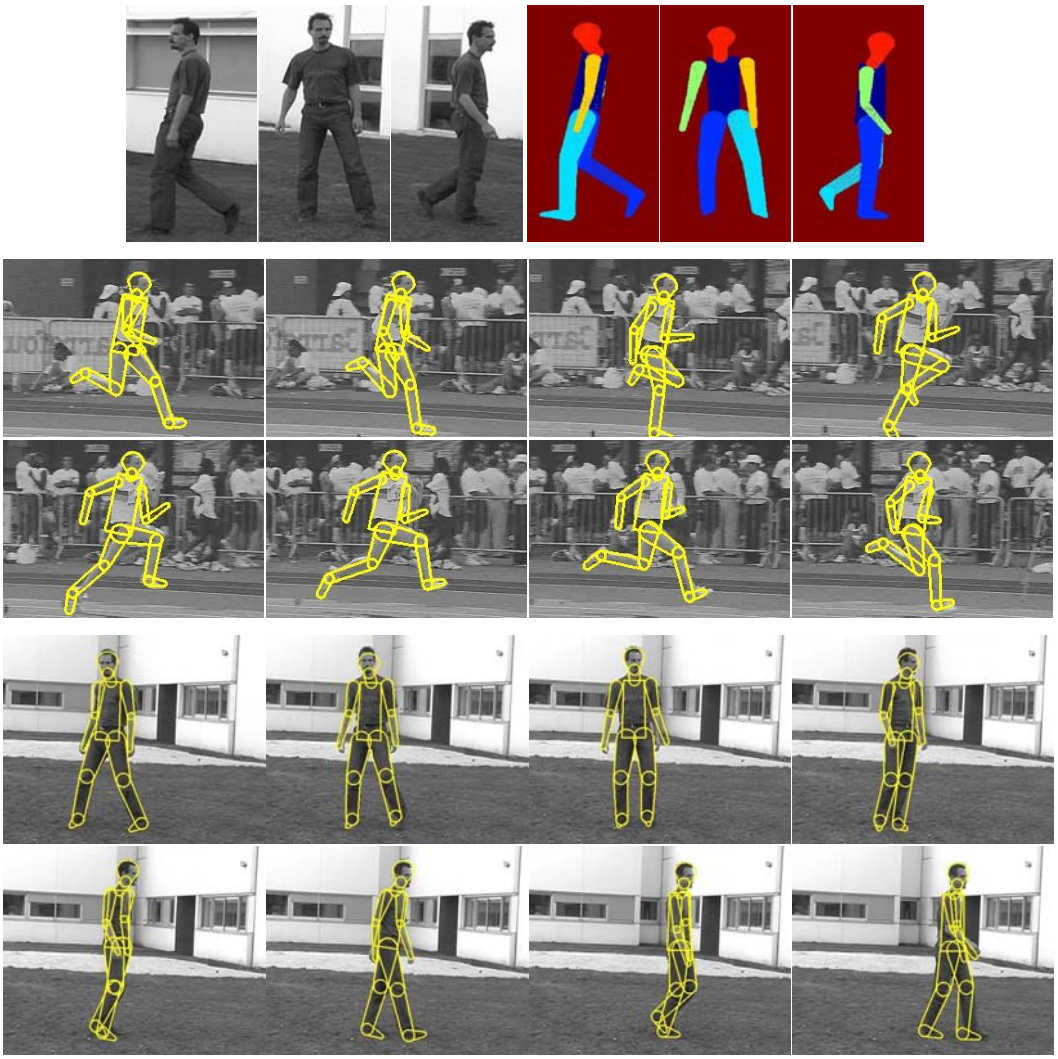
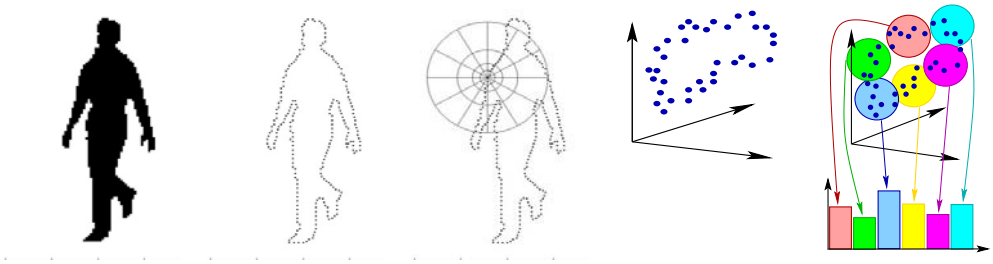


Figure 26: Some examples of 2D human tracking with our learned dynamical model. First row: In preparation for learning, the configuration of the model is marked by hand in each training image. Here we show three images with their corresponding configurations and visibility maps. Rows 2–3: Tracking a running athlete. The model was learned on another athlete but follows this one well except that the left arm was not correctly initialized because it was invisible in the initial image. Last 2 rows: Tracking a turning movement through changes of model aspect from face to side view.

for relatively complex motions such as turns. Figure 26 shows two examples. We are also investigating the use of these kinds of models for classifying different categories of human behaviour.

Learning to reconstruct 3D human motion from silhouettes: The above approaches to human motion capture involve fitting fairly complicated articulated models that can be difficult to initialize and ambiguous to track. The main source of ambiguity is perhaps the fact that they attempt to model



(a) silhouette extraction

(b) sample points around the silhouette

(c) evaluate their shape contexts

(d) distribution of responses in shape context space

(e) vector quantize to get the descriptor

Figure 27: The extraction process for our robust descriptors of silhouette geometry. "Shape context" descriptors are evaluated at regular points around the silhouette, and the distribution of responses in shape context space is vector quantized into 100 bins to produce the final descriptor.

all *possible* human poses. They do not incorporate much prior knowledge that would allow attention to be focused principally on the smaller set of *typical* poses. For this reason they waste a considerable amount of effort searching over poses that are simply (to humans) implausible, and they often mistrack by following such a hypothesis too far. An alternative strategy, which sidesteps both the initialization and the implausibility problems, is to discard the explicit manually-constructed articulated body model, and instead to directly *learn* to estimate pose from a suitable set of low-level descriptors extracted from the image.

We have developed several approaches of this type, that directly regress full 3-D body pose (encoded by joint angle variables) against robust image descriptors that characterize the geometry of the human's image silhouette. The descriptors used — vector quantized versions of the distribution of shape-context responses at boundary points of the silhouette — give good locality (and hence some occlusion resistance) and robustness to the segmentation errors that are so common with real image silhouettes obtained by background subtraction. Figure 27 illustrates the process. Various different regressors have been tested. Linear regression (on our very nonlinear descriptors) already works well, but the best results are with a sparse Bayesian kernel based method, Relevance Vector Regression.

The methods have been tested on both synthetic images of unseen sequences (to allow ground-truthing) and real images. The training data is real human motion capture data for a variety of human motions, together with the corresponding image silhouettes. The use of real motions ensures that the poses that appear are typical for humans.

Our initial method, begun at the end of last year, is based on separate reconstruction in each image [11]. It achieves average 3D joint angle estimation errors of around 6–7° — significantly better than existing learning-based approaches, but not yet satisfying from an applications point of view, especially these estimates include frequent (15%) gross misestimations owing to ambiguities intrinsic to the silhouette based representation. To rectify this, we have developed two additional methods this year. The first incoporates the initial method in a dynamical tracking based framework, giving smooth reconstructions and reasonably satisfying reanimations [13]. Figure 28 (top) shows some results on a real test sequence. The second approach keeps the original static image based framework, but uses

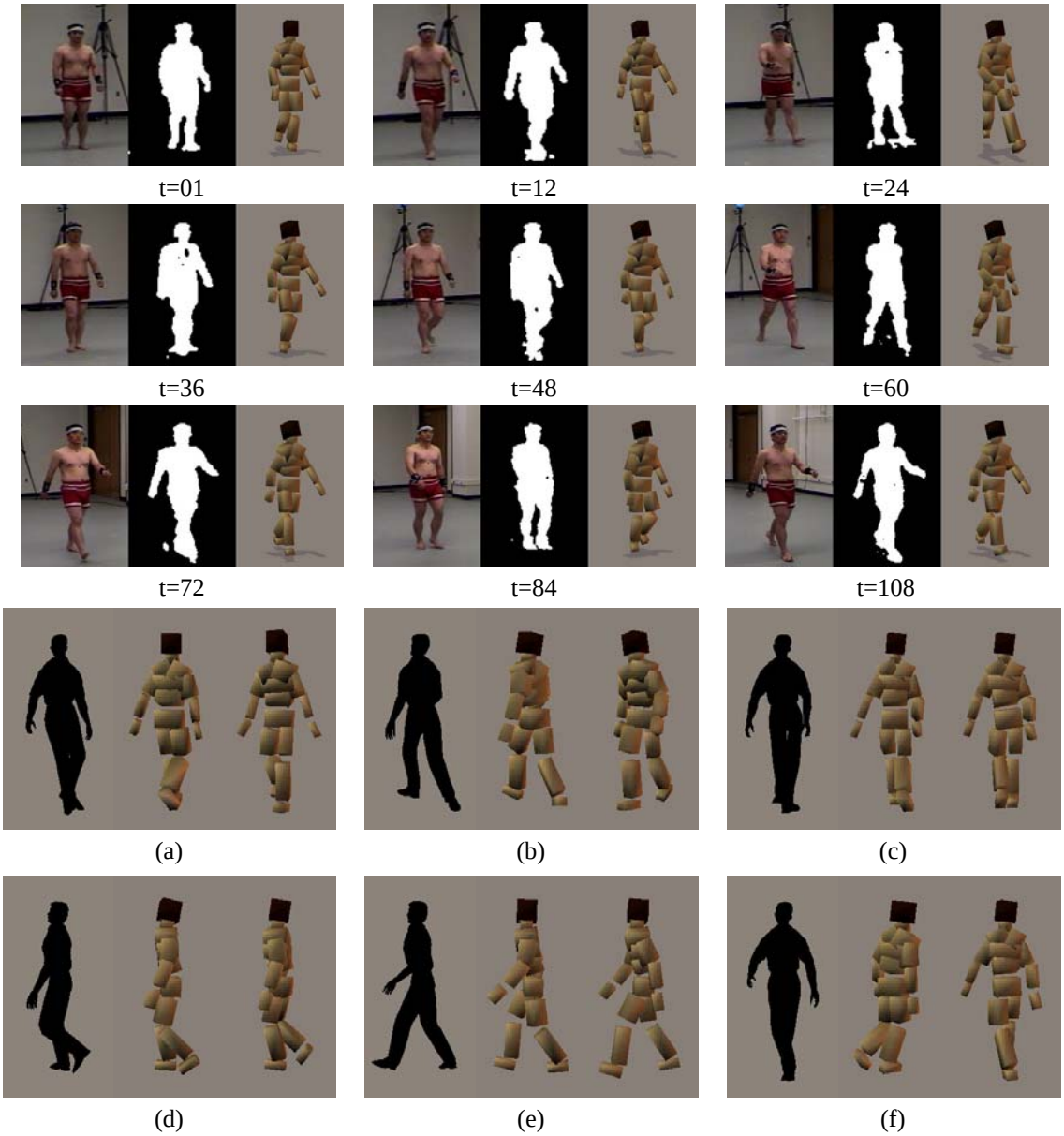


Figure 28: *Top:* 3D poses reconstructed from a test video sequence obtained from *http://mocap.cs.cmu.edu*. In this sequence, the subject walks towards the camera causing a scale change by a factor of two. (The images and silhouettes have been normalized for display). Our representation allows the algorithm to process a silhouette independent of its size or location in the image without disturbing the 3D pose recovery. *Bottom:* Multiple possible 3D pose estimates obtained from individual silhouettes using a mixture of regressors. The two most likely modes of the distribution are shown in each case, and generally capture the two most evident reconstruction possibilities, 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 first solution is a misestimate but feasible solutions are obtained in the other modes.

multi-valued regression to generate several possible solutions from each silhouette, together with posterior probabilities for them to be the correct reconstruction [26]. The solution is often essentially unique, and even when there is ambiguity it is rare to find more than 2–3 solutions. Figure 28 (bottom) shows some sample reconstructions.

7 Contracts and Grants with Industry

7.1 Pandora Studio

Participants: Bill Triggs, Marius Malciu [Pandora Studio].

Our collaboration, begun in June 2003, with the French audiovisual and animation consultancy Pandora Studio based in Paris and Toulouse, continued until June 2004. The object was to develop software for semi-automatic 3D human motion capture from single image streams, for production studio applications (film, TV, games).

7.2 Bertin Technologies

Participants: Frederic Jurie, Cordelia Schmid, Roger Mohr, Eric Nowak.

The aim of the collaboration with Bertin Technologies is to develop algorithms for detecting and recognizing objects in the context of an unmanned infra-red information systems. The challenges are the poor resolution of images, the changeable appearances of objects due to temperature changes (global and local), and the high number of nested object categories. Applications are typically outdoors defense applications in which hidden cameras have been left to detect the presence of military vehicles. Our collaboration is funded by a PhD thesis (CIFRE funds) that started in March 2004. Bertin Technologies is also one of our partners in the Techno-Vision project, see paragraph 8.1.2. We plan to extend our collaboration with Bertin Technologies. We have, for example, answered a call for SCOOP project together, for which we are currently shortlisted. SCOOP is a national project aiming to provide tools for image retrieval and image databases.

7.3 MBDA

Participants: Frederic Jurie, Cordelia Schmid.

We have been collaborating with the Aerospatiale section of MBDA over the past few years. This year we produced a state of the art on object recognition as a part of this. In 2005, we plan to extend the collaboration by starting a PhD thesis or post-doc funded by the company. MBDA is also one of our partners in the Techno-Vision project.

7.4 THALES Optronics

Participants: Frederic Jurie, Diane Larlus.

In 2004, we began a collaboration with THALES Optronics. The aim is to develop algorithms for the detection of objects in aerial images. In particular, it concentrates on the selection of reliable parts

based on the clustering of a large set of features. In 2005, they will finance the DEA of D. Larlus on this subject and they are planing to co-finance a PhD thesis.

8 Other Grants and Activities

8.1 National grants

8.1.1 Ministry grant MoViStaR

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

MoViStaR is a joint national project ("action concertée incitative") under the program "Masses de Données" (Large Quantites of Data). The partners are the INRIA/Gravir project LEAR (C. Schmid, B. Triggs), the INRIA team MISTIS (F. Forbes), the SMS team of the LMC laboratory (S. Girard) and the Heudiasyc laboratory (C. Ambroise). MoViStaR started in September 2003 for three years. It aims at developing techniques for mining visual information from large image collections, to achieve reliable recognition of object categories. In particular, it concentrates on applying and adapting statistical data reduction techniques and on the integration of spatial information.

8.1.2 Techno-Vision

Participants: Frederic Jurie, Roger Mohr, Cordelia Schmid.

In 2004, we have been chosen to lead the national project Techno-Vision ROBIN, which aims at producing datasets, ground truth, competitions and metrics for the evaluation of object recognition algorithms. ROBIN is a two year project 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 categorisation.

The project datasets and metadata will be produced during the first year (2005/2006), and the second year will be devoted to the organisation of the competition (selection of the test images and of the benchmarking procedure).

8.2 European Projects

8.2.1 VIBES

Participants: Cordelia Schmid, Bill Triggs, Ankur Agarwal, Salil Jain, Michael Sdika, Krystian Mikolajczyk [Oxford], Andrew Zisserman [Oxford].

VIBES (Video Browsing, Exploration and Structuring) was a 5th framework FET-Open project. It started in November 2000 and ran for 3 years and 5 months, ending in April 2004. Its main goal was the automatic extraction of object-level representations from video streams. The partners were KTH Stockholm, Sweden (coordinator), LEAR (France), Oxford University (UK), the Katholieke Universiteit Leuven (Belgium), the Ecole Polytechnique Fédérale de Lausanne (Switzerland), and the Weiz-

mann Institute of Science (Israel). VIBES represented an effort of 32 person-years with a total budget of 2.3 MEu, of which 1.7 MEu was funded by the European community. LEAR worked on the video indexing, and human tracking and reconstruction themes of VIBES.

8.2.2 LAVA

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

Learning for Adaptable Visual Assistants (LAVA) is a 5th framework RTD project started in May 2002 for 3 years. It aims at developing 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. LAVA is an interdisciplinary project, involving teams working on machine learning, computer vision, and cognitive modeling and data fusion. The coordinator is Xerox Research Centre Europe (XRCE, Grenoble, France), and the other partners are: 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 LAVA will involve 51 person-years of research effort, for a total budget of 4.3 MEu, including 2.4 MEu of European Union support. LEAR is working mainly on the development of image descriptors for individual images, the interface between vision and learning, and semi-supervised learning.

8.2.3 PASCAL

Participants: Bill Triggs, Cordelia Schmid, Ankur Agarwal, Charles Bouveyron, Guillaume Bouchard, Gyuri Dorko.

PASCAL (Pattern Analysis, Statistical Modelling and Computational Learning) is a Network of Excellence that started in December 2003 for four years, funded by the Multimodal Interfaces theme of the EU 6th framework. The focus is on applying advanced machine learning and statistical pattern recognition techniques to the analysis of various types of sensed data. It unites around 120 European researchers and 100 doctoral students in machine learning, 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 Prof. John Shawe-Taylor of Southampton University. Bill Triggs of LEAR is coordinating the computer vision aspects, the Multimodal Interfaces Thematic Program and various management activities including the overall thematic balance of the network (Balance & Integration Program and Funding Review Program). LEAR and Xerox Research Europe (XRCE) together form one of PASCAL's 14 key sites, focusing on computer vision and natural language processing.

8.2.4 AceMedia

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

AceMedia is a 6th framework Integrated Project that will run for 4 years starting from January 2004. Its goal is 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. LEAR is working mainly on human detection and action recognition in static images and in videos.

8.3 Bilateral relationship

8.3.1 University of Oxford, UK

Participants: Cordelia Schmid, Bill Triggs, Krystian Mikolajczyk [Oxford], Andrew Zisserman [Oxford].

In 2004, the collaboration with the research group of A. Zisserman in the University of Oxford was partially funded by the European project VIBES. The collaboration focuses on invariant local feature detectors and human detection. In 2004, K. Mikolajczyk visited Grenoble for one month and B. Triggs spent three days in Oxford.

8.3.2 University of Illinois at Urbana-Champaign, USA

Participants: Cordelia Schmid, Gyuri Dorko, Svetlana Lazebnik [UIUC], Jean Ponce [UIUC], Fred Rothganger [UIUC].

The research project on 3D object recognition between the research group of J. Ponce and LEAR is funded by a CNRS/UIUC collaboration agreement. In 2004, C. Schmid visited the partner institution for one week. She also participated in Fred Rothganger's PhD committee.

8.3.3 Australian National University and National ICT Australia

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

This collaboration currently 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 2 weeks in March 2004 to work on local feature methods for location recognition.

9 Dissemination

9.1 Leadership within scientific community

- Organization of conference and workshops:
  - Program chair for IEEE Conference on Computer Vision 2005 (C. Schmid)

  - Co-Organizer and program chair of the International Workshop on “Designing Tomorrow’s Category-Level 3D Object Recognition Systems”, Taomina, Sicily, Italy, September 2004 (C. Schmid, B. Triggs)
  - Chair of PASCAL workshop on Pattern Recognition and Machine Learning in Computer Vision, Grenoble, May 2004 (B. Triggs)
  - Workshops chair of the 10th IEEE International Conference on Computer Vision 2005 (B. Triggs)

- Editorial board:
  - International Journal of Computer Vision (C. Schmid)
  - IEEE Transactions on Pattern Analysis and Machine Intelligence (C. Schmid)
  - Machine Vision and Applications (R. Mohr)
  - Techniques et Sciences Informatiques (F. Jurie)

- Area chair :
  - CVPR’04 (C. Schmid)
  - ECCV’04 (C. Schmid)
  - RFIA’04 (C. Schmid)
  - ICCV’05 (C. Schmid, B. Triggs)

- Program committee :
  - RFIA’04 (F. Jurie)
  - ECCV’04 (F. Jurie, B. Triggs)
  - CVPR’04 (F. Jurie, B. Triggs)
  - ICML’04 (B. Triggs)
  - NIPS’04 (C. Schmid, B. Triggs)
  - CVPR’05 (F. Jurie, B. Triggs)

- Other
  - C. Schmid is a member of the “INRIA commission d’évaluation” and of the “INRIA RA comité des emplois scientifiques”
  - F. Jurie is vice-head of AFRIF (French section of IAPR)
  - B. Triggs manages the overall Balance & Integration of the EU Network of Excellence PASCAL.
  - B. Triggs served as an expert at the EU consultation meeting on MultiModal Interfaces, Luxembourg, May 2004.

- Prizes :

– Guillaume Bouchard of LEAR and SELECT received the prize for the best student paper at the 14th Congrès de la Maîtrise des Risques et de la Sûreté de Fonctionnement (French national Conference on Risk Management and Reliability) for his paper *Réactualisation bayésienne d'un modèle de dégradation en fonction du retour d'expérience* (*Bayesian Updating of a Failure Model based on Experimental Data*) [15].

9.2 Teaching

- Matching and Recognition, DEA IVR, INPG, 12 h (C. Schmid, F. Jurie)
- Multi-media database, 3rd year ENSIMAG, INPG, 11 h (F. Jurie)

9.3 Invited presentations

- C. Schmid. *Building Local Part Models for Category-Level Recognition*. Cognitive Computer Vision Colloquium, Prague, January 2004.
- B. Triggs. *Learning to Recover 3D Human Pose from Silhouettes.* Australian National University, Canberra, Australia, March 2004.
- B. Triggs. *Learning to Recover 3D Human Pose from Silhouettes.* Learning Workshop, Snowbird, Utah, April 2004.
- C. Schmid. *Building Local Part Models for Category Recognition.* Vision Seminar at Berkeley University, April 2004.
- C. Schmid. *Affine-Invariant Local Features and Applications to Recognition*. PASCAL Workshop on Pattern Recognition and Machine Learning, Grenoble, May 2004.
- R. Mohr. *Image plus Data Base differs from Image Data Base*. Sigmod Workshop on Computer Vision meets Data Base, Paris, June 2004.
- C. Schmid. *Comparison of Affine Covariant Detectors and Descriptors*. International Workshop on Object Recognition, Taomina, Sicily, Italy, October 2004.
- B. Triggs. *A Hierarchical Part-Based Model for Visual Object Categorization*. International Workshop on Object Recognition, Taomina, Sicily, Italy, October 2004.
- B. Triggs. *Learning Based Methods for Human Motion Capture.* University of Oxford, November 2004.
- C. Schmid. *Building Local Part Models for Category-Level Recognition*. Seminar at Max Planck Institut Tübingen, December 2004.

10 Bibliography

Articles in referred journals and book chapters

[1] Y. DUFOURNAUD, C. SCHMID, R. HORAUD, "Image matching with scale adjustment", *Computer Vision and Image Understanding 93*, 2, 2004, p. 175–194, http://lear.inrialpes.fr/pubs/2004/DSH04.

[2] S. LAZEBNIK, C. SCHMID, J. PONCE, "A sparse texture representation using local affine regions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, accepted, http://lear.inrialpes.fr/pubs/2004/LSP04a.

[3] K. MIKOLAJCZYK, C. SCHMID, "A performance evaluation of local descriptors", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, accepted suject to minor revision, http://lear.inrialpes.fr/pubs/2004/MS04a.

[4] K. MIKOLAJCZYK, C. SCHMID, "Scale and affine invariant interest point detectors", *International Journal of Computer Vision 60*, 1, 2004, p. 63–86, http://lear.inrialpes.fr/pubs/2004/MS04.

[5] 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*, 2004, accepted suject to minor revision, http://lear.inrialpes.fr/pubs/2004/RLSP04.

[6] 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 Publishing Co., 2004, to appear, http://lear.inrialpes.fr/pubs/2004/SDLMP04.

[7] C. SCHMID, "Recognition with local photometric invariants", *in : Trends and Advances in Content-Based Image and Video Retrieval*, L. Shapiro, H. Kriegel, and R. Veltkamp (editors), Springer, 2004, to appear, http://lear.inrialpes.fr/pubs/2004/Sch04a.

[8] C. SCHMID, "Weakly supervised learning of visual models and its application to content-based retrieval", *International Journal of Computer Vision 56*, 1, 2004, p. 7–16, http://lear.inrialpes.fr/pubs/2004/Sch04.

[9] C. SMINCHISESCU, B. TRIGGS, "Hyperdynamic Sampling", *Journal of Image & Vision Computing*, 2004, special issue on ECCV'02 papers, to appear in early 2005, http://lear.inrialpes.fr/pubs/2004/ST04a.

[10] 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.

Publications in Conferences and Workshops

[11] 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.

[12] A. AGARWAL, B. TRIGGS, "Learning to Recover 3D Human Pose from Silhouettes", *in : Learning 2004 - Abstracts of the 2004 Snowbird Learning Workshop*, Y. LeCun, Y. Bengio (editors), Computational and Biological Learning Society, April 2004, http://lear.inrialpes.fr/pubs/2004/AT04c.

[13] 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.

[14] A. AGARWAL, B. TRIGGS, "Tracking Articulated Motion using a Mixture of Autoregressive Models", *in : 8th European Conference on Computer Vision*, p. III 54–65, Prague, Czech Republic, May 2004, http://lear.inrialpes.fr/pubs/2004/AT04a.

[15] G. BOUCHARD, G. CELEUX, F. BILLY, F. JOSSE, "Réactualisation bayésienne d'un modèle de dégradation en fonction du retour d'expérience", *in : Colloque de Maîtrise des risques et de Sûreté de Fonctionnement (Lambda-Mu)*, Bourges, France, 2004.

[16] G. BOUCHARD, B. TRIGGS, "The Tradeoff Between Generative and Discriminative Classifiers", *in : IASC International Symposium on Computational Statistics (COMPSTAT)*, p. 721–728, Prague, August 2004, http://lear.inrialpes.fr/pubs/2004/BT04.

[17] C. BOUVEYRON, S. GIRARD, C. SCHMID, "Dimension Reduction and Classification Methods for Object Recognition in Vision", *in : 5th French-Danish Workshop on Spatial Statistics and Image Analysis in Biology*, p. 109–113, May 2004, http://lear.inrialpes.fr/pubs/2004/BGS04.

[18] 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, http://lear.inrialpes.fr/pubs/2004/JS04.

[19] S. LAZEBNIK, C. SCHMID, J. PONCE, "Learning local affine-invariant part models for object class recognition", *in : Learning 2004 - Abstracts of the 2004 Snowbird Learning Workshop*, Y. LeCun, Y. Bengio (editors), Computational and Biological Learning Society, April 2004, http://lear.inrialpes.fr/pubs/2004/LSP04b.

[20] S. LAZEBNIK, C. SCHMID, J. PONCE, "Semi-local Affine Parts for Object Recognition", *in : British Machine Vision Conference, volume 2*, p. 779–788, 2004, http://lear.inrialpes.fr/pubs/2004/LSP04.

[21] K. MIKOLAJCZYK, C. SCHMID, A. ZISSERMAN, "Human detection based on a probabilistic assembly of robust part detectors", *in : 8th European Conference on Computer Vision, I*, p. 69–81, Prague, Czech Republic, 2004, http://lear.inrialpes.fr/pubs/2004/MSZ04.

[22] K. MIKOLAJCZYK, C. SCHMID, "Comparison of affine-invariant local detectors and descriptors", *in : 12th European Signal Processing Conference*, Vienna, Austria, 2004, http://lear.inrialpes.fr/pubs/2004/MS04b.

[23] J. PONCE, S. LAZEBNIK, F. ROTHGANGER, C. SCHMID, "Towards true 3D object recognition", *in : Reconnaissance des Formes et Intelligence Artificielle*, 2004, http://lear.inrialpes.fr/pubs/2004/PLRS04.

[24] F. ROTHGANGER, S. LAZEBNIK, C. SCHMID, J. PONCE, "Segmenting, modeling and matching video clips containing multiple moving objects", *in : IEEE Conference on Computer Vision and Pattern Recognition, 2*, p. 914–921, Washington DC, USA, 2004, http://lear.inrialpes.fr/pubs/2004/RLSP04a.

[25] B. TRIGGS, "Detecting Keypoints with Stable Position, Orientation and Scale under Illumination Changes", *in : 8th European Conference on Computer Vision*, p. IV 100–113, Prague, Czech Republic, May 2004, http://lear.inrialpes.fr/pubs/2004/Tri04.

Internal Reports

[26] A. AGARWAL, B. TRIGGS, "Learning Methods for Recoverng 3D Human Pose from Monocular Images", *Research Report number 5333*, INRIA Rhone Alpes, 655 ZIRST, Avenue de l'Europe, 38330 Montbonnot, France, October 2004, http://lear.inrialpes.fr/pubs/2004/AT04d.

Miscellaneous

[27] J. BLANCHET, *Modèles Markoviens pour l'organisation spatiale des descripteurs en reconnaissance de textures*, Mémoire, DEA de Mathématiques, Université Toulouse III, Juin 2004.

[28] A. BUGEAU, *Attention visuelle multi-échelle*, Mémoire, DEA Image Vision Robotique, Institut National Polytechnique de Grenoble, July 2004, http://lear.inrialpes.fr/pubs/2004/Bug04.

[29] G. DORKÓ, C. SCHMID, "Object class recognition using discriminative local features", Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, http://lear.inrialpes.fr/pubs/2004/DS04.

[30] K. MIKOLAJCZYK, T. TUYTELAARS, C. SCHMID, A. ZISSERMAN, J. MATAS, F. SCHAFFALITZKY, T. KADIR, L. V. GOOL, "A comparison of affine region detectors", Submitted to International Journal of Computer Vision, 2004, http://lear.inrialpes.fr/pubs/2004/MTSZMSKG04.

[31] B. TRIGGS, "Boundary Conditions for Young - van Vliet Recursive Filtering", Submitted to IEEE Transactions on Image Processing, June 2004, http://lear.inrialpes.fr/pubs/2004/Tri04a.