INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team WILLOW

Models of Visual Object Recognition and Scene Understanding

Paris - Rocquencourt

THEME COG

Table of contents

1. Team

Willow is a common project with l'Ecole Normale Supérieure de Paris. The team has been created on January the 1st, 2007 and became an INRIA project on June the 27th, 2007.

Head of project-team

Jean Ponce [Professor in the Département d'Informatique of École Normale Supérieure (ENS), and adjunct professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign (UIUC), HdR]

Vice-head of project-team

Andrew Zisserman [Professor in the Engineering Department of the University of Oxford, and part-time professor at ENS funded by an EADS industrial chair, HdR]

Administrative assistant

Nathalie Abiola

Research scientists

Jean-Yves Audibert [Chercheur at the Centre d'Enseignement et de Recherche en Technologies de l'Information et Systèmes (CERTIS) of the École Nationale des Ponts et Chaussées (ENPC)]
Francis Bach ["Détaché" at INRIA from the Corps des Mines]
Jean-Philippe Pons [Chercheur at CERTIS - ENPC]
Florent Ségonne [Chercheur at CERTIS - ENPC]
Josef Sivic [will start as an INRIA research scientist (CR2) in January 2008]

Post-doctoral fellow

Bryan Russell

PhD students

Eshan Aganj
Y-Lan Boureau
Patrick Etyngier
Akash Kushal
Patrick Labatut
Julien Mairal
Oliver Whyte

Student interns

Jérôme Courchay
Mariano Tepper

2. Overall Objectives

2.1. Overall Objectives

Object recognition —or, in a broader sense, scene understanding— is the ultimate scientific challenge of computer vision: After 40 years of research, robustly identifying the familiar objects (chair, person, pet), scene categories (beach, forest, office), and activity patterns (conversation, dance, picnic) depicted in family pictures, news segments, or feature films is still far beyond the capabilities of today's vision systems. On the other hand, truly successful object recognition and scene understanding technology will have a broad impact in application domains as varied as defense, entertainment, health care, human-computer interaction, image retrieval and data mining, industrial and personal robotics, manufacturing, scientific image analysis, surveillance and security, and transportation.

Despite the limitations of today's scene understanding technology, tremendous progress has been accomplished in the past ten years, due in part to the formulation of object recognition as a statistical pattern matching problem. The emphasis is in general on the features defining the patterns and on the algorithms used to learn and recognize them, rather than on the representation of object, scene, and activity categories, or the integrated interpretation of the various scene elements. WILLOW complements this approach with an ambitious research program explicitly addressing the representational issues involved in object recognition and, more generally, scene understanding.

Concretely, our objective is to develop geometric, physical, and statistical models for all components of the image interpretation process, including illumination, materials, objects, scenes, and human activities. These models will be used to tackle fundamental scientific challenges such as three-dimensional (3D) object and scene modeling, analysis, and retrieval; human activity capture and classification; and category-level object and scene recognition. They will also support applications with high scientific, societal, and/or economic impact in domains such as quantitative image analysis in science and humanities; film post-production and special effects; and video annotation, interpretation, and retrieval. Machine learning is a key part of our effort, with a balance of practical work in support of computer vision application, methodological research aimed at developing effective algorithms and architectures, and foundational work in learning theory.

WILLOW was created in 2007: It was recognized as an INRIA team in January 2007, and as an official project-team in June 2007. WILLOW was originally conceived as a joint venture between Ecole Normale Supérieure (ENS), INRIA Paris Rocquencourt (COG B research theme), and Ecole Nationales des Ponts et Chaussées (ENPC). Two of its original members, Jean-Philippe Pons and Florent Ségonne, are leaving our team, due to an internal reorganization of CERTIS, the ENPC computer science department, that is causing the departure of most of its members from the INRIA project-teams they were associated with. Thus WILLOW should be considered, from 2008 on, as a joint ENS/INRIA project-team only, although Jean-Yves Audibert, the third ENPC member of WILLOW, remains with us. The departure of Pons and Ségonne is somewhat compensated by a very successful recruiting season, since two new researchers have replaced them, Francis Bach, who joined us in September 2007, on leave ("détachement") for five years from the "Corps des Mines", and Josef Sivic, who was hired as research scientist ("chargé de recherches") and will join us in January 2007. In addition, two PhD students from the "Corps des Télécom", Y-Lan Boureau and Julien Mairal, and a post-doc from MIT, Bryan Russell, have joined WILLOW in 2007.

3. Scientific Foundations

3.1. 3D object and scene modeling, analysis, and retrieval

This part of our research focuses on geometric models of specific 3D objects at the local (differential) and global levels, and physical and statistical models of materials and illumination patterns. Our past work in this area includes research aimed at recognizing rigid 3D objects in cluttered photographs taken from arbitrary viewpoints (Rothganger *et al.*, 2006), and segmenting video sequences into parts corresponding to rigid scene components before recognizing these in new video clips (Rothganger *et al.*, 2007). Our current research focuses on the acquisition of detailed object and scene models from multiple images and video streams:

3.1.1. High-fidelity image-based object and scene modeling.

As further detailed in Section 6.1, we have recently developed several algorithms for multi-view stereopsis [23], [26] that have proven remarkably effective at recovering intricate details and thin features of compact objects and capturing the overall structure of large-scale, cluttered scenes. is available for free for academics, and licensing negociations with several companies are under way. We have also recently developed a new calibration algorithm that uses rough multi-view reconstructions to obtained extremely accurate intrinsic and extrinsic camera parameters.

3.1.2. Video-based modeling of deformable surfaces.

As discussed in Section 6.2, we have also generalized our work on multi-view stereopsis to the dynamic analysis of video streams that depict objects with deformable surfaces, for example walking persons, human faces, and folding cloth. These approaches exploit the spatio-temporal consistency of image sequences [14] and locally rigid but globally nonrigid models of surface motion to accurately capture the deforming shape of the observed surfaces.

3.2. Category-level object and scene recognition

The objective in this core part of our research is to learn and recognize quickly and accurately thousands of visual categories, including materials, objects, scenes, and broad classes of temporal events, such as patterns of human activities in picnics, conversations, etc. The current paradigm in the vision community is to model/learn one object category (read 2D aspect) at a time. If we are to achieve our goal, we have to break away from this paradigm, and develop models that account for the tremendous variability in object and scene appearance due to texture, material, viewpoint, and illumination changes within each object category, as well as the complex and evolving relationships between scene elements during the course of normal human activities. Our current work focuses on the following problems:

3.2.1. Learning image and object models.

We have introduced two *sparse* models of local image features that are adapted to object recognition tasks, and can effectively be learned from training data. The first one, developed in collaboration with Yann LeCun and his students at NYU, has given good results in handwritten digit recognition tasks [27]. The second model, developed in collaboration with Guillermo Sapiro at the University of Minnesota, generalizes to discriminative tasks an approach originally developed for image restoration [10], and it has been shown to be effective in texture segmentation and feature selection tasks. We have also introduced a novel model of higher-level image *attributes* and used it in image search tasks. Finally, we have developed a new model of object categories that explicitly captures image variations due to shape and viewpoint changes within a category, and demonstrated its ability to detect objects such as cars in images despite such changes [25].

3.2.2. Image segmentation.

Segmentation is one of the most difficult problems in computer vision (see [15] for a unified view of several classical algorithms). As a purely bottom-up process of distinguishing foreground from background image regions without any a priori information, it is also ill posed. We have considered instead in our recent work two instances of this problem where top-down information is used to make it well posed. In [20], [19], this takes the form of non-linear shape priors learned from sample shapes. We have recently proposed a supervised energy-based formulation that uses global image features to iteratively improve an initial segmentation based on learned local classifiers.

3.3. Machine learning

3.3.1. Machine learning for computer vision.

A large portion of research in computer vision involves increasingly more refined machine learning techniques. Significant success has been obtained by the direct use of off-the-shelf techniques, such as kernel methods (support vector machines for example) and probabilistic graphical models. However, in order to achieve the level of performance that we aim for, a more careful integration of machine learning and computer vision algorithmic and theoretical frameworks is needed. A major part of our machine learning effort is dedicated to this integration, through: (a) applying the *transductive learning* framework to exploit the simultaneous availability of training and test data in semi-interactive segmentation tasks, (b) using specific kernel designs for images, allowing the natural topological and geometrical structure of images to be taken into account, thus allowing a considerable reduction in the number of labelled examples (Harchaoui and Bach, 2007), and (c) developing efficient approximate inference algorithms for graphical models with geometric constraints, allowing a more faithful probabilistic model for scene analysis.

3.3.2. Effective learning algorithms and architectures.

Probabilistic graphical models provide a very flexible and powerful framework for capturing statistical dependencies in complex, multivariate data. The main current methodological bottleneck in their application is the computational complexity of the inference. We are currently investigating the links between the various state-of-the-art techniques for approximate inferences (variational methods, simulation methods and graph cuts). Another key part of our algorithmic research is dedicated to semi-supervised and active learning: in many domains, such as vision or bioinformatics, large databases are available but only with a few labelled examples. In this setting, semi-supervised learning aims at using the unlabelled examples in order to improve the prediction performance, while active learning aims at optimizing the selection of examples to label in order to maximize the final predictive performance. Although many algorithms have been proposed, few of them have theoretical and practical guarantees regarding their predictive performances, and our research effort will be dedicated to the design of robust and efficient algorithms for active and semi-supervised learning, following our earlier work (Bach, 2006). Finally, the computational complexity of very simple computer vision tasks (e.g. object matching) is such that it is often impossible to use these tasks to extract knowledge from large image database or video sequences. We intend to address the problem of efficient use of data and computational resources. In particular, we will develop our research on the exploration-exploitation dilemma (see Audibert, Munos and Szepesvari, 2007) and focus on hierarchical structures.

3.3.3. Learning theory.

We aim at providing a better understanding of the fundamental ideas underlying efficient learning algorithms. To understand well popular methods is often a key step in order to refine and generalize these methods, and also to design new learning algorithms. Apart from the computational complexity mentioned before, the common features encountered when using learning techniques in computer vision are (i) high dimensionality and (ii) complexity of the modelization. To avoid the curse of dimensionality, we intend to search for sparse representations of the prediction function. Sparsity inducing norms are raising increased interest in the statistics and learning theory communities; regularizing learning problems using such norms leads to both sparse predictors and good generalization performances. Recent research has thoroughly looked at the behavior of regularization by the 1-norm (sum of absolute values), and there is currently a strong effort in extending those results to other more complex settings (e.g., Bach, 2007). To get round the modelization problem, a standard way is to consider embedded models of increasing complexity. We intend to develop adaptive learning procedures predicting as well as the best model in the nested family.

3.4. Human activity capture and classification

We have left this essential area of our planned activities for last in this section since it will only really take off in 2008. From a scientific point of view, visual action understanding is a computer vision problem that has received little attention so far outside of extremely specific contexts such as surveillance or sports. Current approaches to the visual interpretation of human activities are designed for a limited range of operating conditions, such as static cameras, fixed scenes, or restricted actions. The objective of this part of our project is to attack the much more challenging problem of understanding actions and interactions in unconstrained video depicting everyday human activities such as in sitcoms, feature films, or news segments. The recent emergence of automated annotation tools for this type of video data (Everingham, Sivic, Zisserman, 2006; Laptev and Pérez, 2006) means that massive amounts of labelled data for training and recognizing action models will at long last be available. This is a fundamental part of our planned research effort, and it will be boosted in 2008 by the arrival of Josef Sivic, whose main area of research has been in video retrieval and indexing, and the start of a collaborative effort on action recognition involving WILLOW, two other INRIA project-teams, LEAR and VISTA, the French “Institut National de l’Audiovisuel” (INA), and Microsoft researchers in Cambridge and elsewhere under the umbrella of an e-science project at the joint MSR-INRIA laboratory.

4. Application Domains

4.1. Introduction

We believe that foundational modeling work should be grounded in applications. This includes (but is not restricted to) the following high-impact domains.

4.2. Quantitative image analysis in science and humanities

We plan to apply our 3D object and scene modeling and analysis technology to image-based modeling of human skeletons and artifacts in anthropology, and large-scale site indexing, modeling, and retrieval in archaeology and cultural heritage preservation. Most existing work in this domain concentrates on image-based rendering—that is, the synthesis of good-looking pictures of artifacts and digs. We plan to focus instead on quantitative applications. A first effort in this area has been a collaboration with the Getty Conservation Institute in Los Angeles, aimed at the quantitative analysis of environmental effects on the hieroglyphic stairway at the Copan Maya site in Honduras. We are now pursuing a larger-scale project involving the archaeology laboratory at ENS and focusing on image-based artifact modeling and decorative pattern retrieval in Pompeii. This new effort is part of the MSR-INRIA project mentioned earlier and that will be discussed further later in this report.

4.3. Film Post-Production and Special Effects

We will apply our 3D object and scene modeling and analysis technology, as well as our human activity capture and classification work to problems such as digital prop and actor capture and tracking, inpainting, and illumination and shadowing. A particularly challenging problem with tremendous applications in film post-production is image-based facial motion capture. This task is made difficult by the (relative) lack of texture and the subtle motions of human faces. We are pursuing these and other applications to post-production and special effects through existing collaborations with Industrial Light and Magic (ILM), the special effects company behind Star Wars and dozens of other Hollywood films.

4.4. Video Annotation, Interpretation, and Retrieval

Both specific and category-level object and scene recognition can be used to annotate, augment, index, and retrieve video segments in the audiovisual domain. The Video Google system developed by Sivic and Zisserman (2005) for retrieving shots containing specific objects is an early success in that area. A sample application, suggested by discussions with Institut National de l'Audiovisuel (INA) staff, is to match set photographs with actual shots in film and video archives, despite the fact that detailed timetables and/or annotations are typically not available for either medium. Automatically annotating the shots is of course also relevant for archives that may record hundreds of thousands of hours of video. Some of these applications will be pursued in our MSR-INRIA project, in which INA is one of our partners.

5. Software

5.1. PMVS

Our multi-view stereopsis PMVS software (http://www-cvr.ai.uiuc.edu/~yfurukaw/research/pmvs/index.html) developed in collaboration with Y. Furukawa at the University of Illinois at Urbana-Champaign [23] is publicly available for academics, and licensing negociations with several companies are under way.

5.2. Structure-from-motion and auto-calibration software

This software was developed by an MVA intern, J. Courchay to complement PMVS and allow the acquisition of accurate object models without the use of cumbersome calibration charts. As this software matures, we intend to make it available to the computer vision community at large.

5.3. Accurate calibration software

Bundled with the two software packages aboves, this programe, developed once again in collaboration with Y. Furukawa at UIUC, forms, a complete package for high-accuracy camera calibration and object and scene modeling. Again, we plan to eventually make this software freely available to academics.

5.4. Visual erosion assessment software

This software was developed by another MVA intern, Mariano Tepper. It is aimed at the quantitative analysis of environmental effects on the hieroglyphic stairway at the Copan Maya site in Honduras, and will shortly be delivered to the Getty Conservation Institute.

6. New Results

6.1. High-fidelity image- and video-based modeling

6.1.1. Accurate, dense, and robust multi-view stereopsis (J. Ponce, joint work with Y. Furukawa, UIUC)

We propose in [23] a novel algorithm for calibrated multi-view stereopsis that outputs a (quasi) dense set of rectangular patches covering the surfaces visible in the input images. This algorithm does not require any initialization in the form of a bounding volume, and it detects and discards automatically outliers and obstacles. It does not perform any smoothing across nearby features, yet is currently the top performer in terms of both coverage and accuracy for four of the six benchmark datasets presented in Seitz *et al.* (2006). The keys to its performance are effective techniques for enforcing local photometric consistency and global visibility constraints. Stereopsis is implemented as a *match, expand, and filter* procedure, starting from a sparse set of matched keypoints, and repeatedly expanding these to nearby pixel correspondences before using visibility constraints to filter away false matches. A simple but effective method for turning the resulting patch model into a mesh appropriate for image-based modeling is also presented. The proposed approach is demonstrated on various datasets including objects with fine surface details, deep concavities, and thin structures, outdoor scenes observed from a restricted set of viewpoints, and "crowded" scenes where moving obstacles appear in different places in multiple images of a static structure of interest (Figure 1).

Figure 1. Sample reconstructions using the PMVS software. In each case, one of the input image is shown, along with views of texture-mapped reconstructed patches and shaded polygonal surfaces.

6.1.2. Multi-view reconstruction of large-scale scenes (P. Labatut and J.-P. Pons, joint work with R. Keriven, ENPC)

Most existing multi-view stereovision approaches require some knowledge of the scene extent and often even of its approximate geometry (e.g., a visual hull). This makes these approaches mainly suited to compact objects admitting a tight enclosing box, imaged on a simple or a known background. We have designed an approach focusing on large-scale cluttered scenes under uncontrolled imaging conditions [26] It first generates a quasi-dense 3D point cloud of the scene by matching keypoints across images in a lenient manner, thus possibly retaining many false matches. Then it builds an adaptive tetrahedral decomposition of space by computing the 3D Delaunay triangulation of the 3D point set. Finally, it reconstructs the scene by labeling Delaunay tetrahedra as empty or occupied, thus generating a triangular mesh of the scene. A globally optimal label assignment, as regards photo-consistency of the output mesh and compatibility with the visibility of keypoints in input images, is efficiently found as a minimum cut solution in a graph.

6.1.3. Accurate camera calibration from multi-view stereo and bundle adjustment (J. Ponce, joint work with Y. Furukawa, UIUC).

The advent of high-resolution digital cameras and sophisticated multi-view stereo algorithms such as those discussed above offers the promises of unprecedented geometric fidelity in image-based modeling tasks, but it also puts unprecedented demands on camera calibration to fulfill these promises. We propose a novel approach to camera calibration where top-down information from rough camera parameter estimates and the output of our PMVS multi-view-stereo system on scaled-down input images are used to effectively guide the search for additional image correspondences and significantly improve camera calibration parameters using the bundle adjustment algorithm of Lourakis and Argyros. The proposed method has been tested on several real datasets—including objects without salient features for which image correspondences cannot be found in a purely bottom-up fashion, and image-based modeling tasks—including the construction of visual hulls where thin structures are lost without our calibration procedure (Figure 2).

6.2. Video-based modeling of deformable surfaces

6.2.1. Spatio-temporal shape from silhouette (J.-P. Pons and F. Ségonne, joint work with E. Aganj and R. Keriven, ENPC)

Shape from silhouette is a popular class of methods for solving the multi-view reconstruction problem in an approximate but efficient and robust manner. Generally, it consists in computing the visual hull, which is the maximal volume consistent with a given set of silhouettes. While many authors have focused on computing the visual hull in the case of static images, leading to several established techniques, very little work has dealt with the case of dynamic scenes captured by multiple video sequences, from an actual spatio-temporal perspective, i.e. by going beyond independent frame-by-frame computations. We have designed a novel method for computing a four-dimensional (4D) representation of the spatio-temporal visual hull of a dynamic scene, based on an extension of a recent provably correct Delaunay meshing algorithm [14]. By considering time as an additional dimension, our approach exploits seamlessly the time coherence between different frames to produce a compact and high-quality 4D mesh representation of the visual hull. The 3D visual hull at a given time instant is easily obtained by intersecting this 4D mesh with a temporal plane, thus enabling interpolation of objects shape between consecutive frames (Figure 3). In addition, our approach offers easy and extensive control over the size and quality of the output mesh as well as over its associated reprojection error.

6.2.2. Dense 3D motion capture from synchronized video streams (J. Ponce, joint work with Y. Furukawa, UIUC).

We propose a novel approach to nonrigid, markerless motion capture from synchronized video streams acquired by calibrated cameras. The instantaneous geometry of the observed scene is represented by a polyhedral mesh with fixed topology. The initial mesh is constructed in the first frame using our PMVS

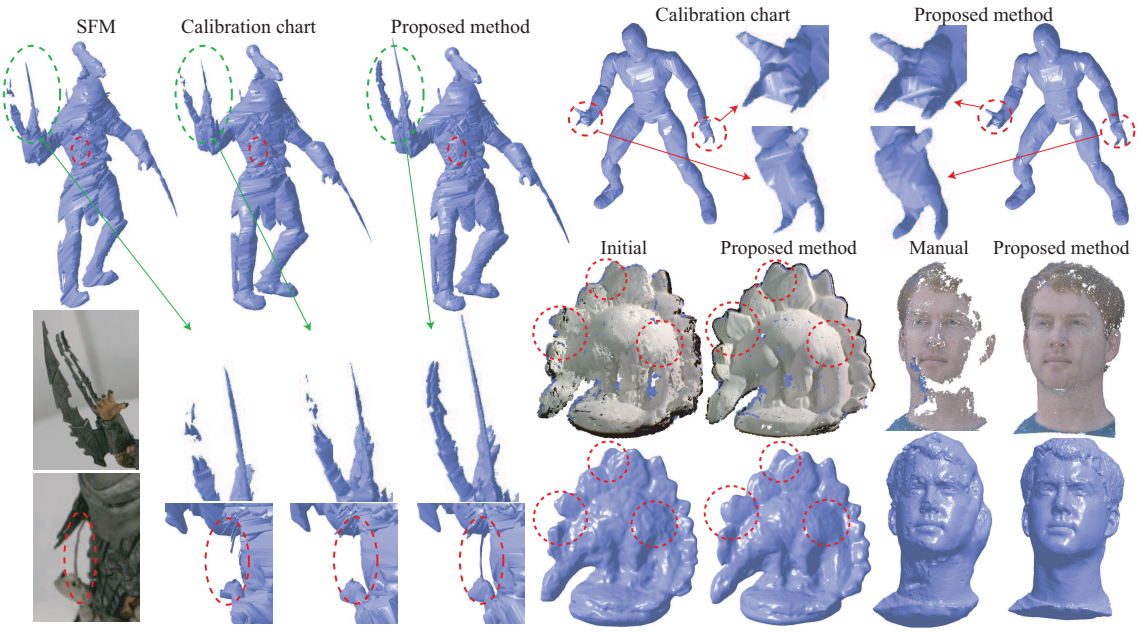

Figure 2. Visual hull models are used to assess the accuracy of camera parameters for spiderman and predator models. Intricate structures are reconstructed only from the camera parameters refined by the proposed method. For dino and face models, a set of patches reconstructed by PMVS and a 3D mesh model extracted from these patches are used for the assessment.

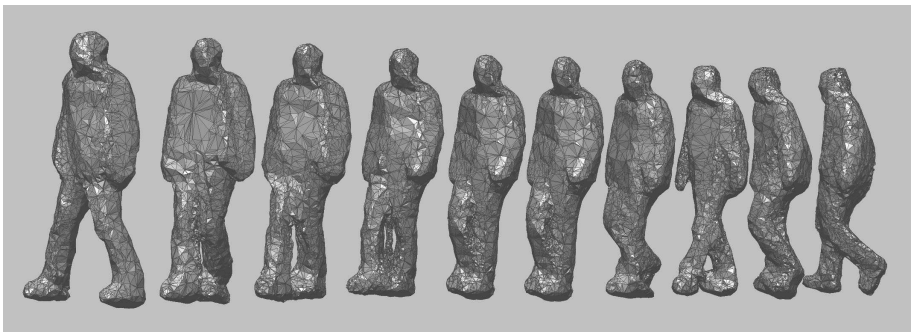

Figure 3. Sample frames of the spatio-temporal visual hull of a walking person.

software for multi-view stereo. Its deformation is captured by tracking its vertices over time, using two optimization processes at each frame: a local one using a rigid motion model in the neighborhood of each vertex, and a global one using a regularized nonrigid model for the whole mesh. Qualitative and quantitative experiments using six real datasets show that our algorithm effectively handles complex nonrigid motions and severe occlusions (Figure 4).

Figure 4. In our approach to motion capture, a polyhedral mesh deforms as its vertices are continuously tracked under locally rigid and globally nonrigid motion models. This is illustrated here with a mesh extracted from real data consisting of 8 synchronized video streams 155 frames long (White et al., 2007). The mesh is shown from two different viewpoints in states 30 frames apart, along with the trajectories of a subset of its vertices (the translational motion is exaggerated for better visualization).

6.3. Learning image and object models

6.3.1. Uncovering higher order correlation in images (Y.-L. Boureau, joint work with M. Ranzato and Y. LeCun, NYU)

Pixels in an image often contain lots of correlations that drastically reduce the apparent dimensionality to a much smaller one. Uncovering these correlations allows to form better representations of images, but can be tricky when it comes to higher order correlations: for instance, it is easy to encode that two neighboring pixels usually behave similarly, forming edges; but encoding angles is wasteful when limiting oneself to pixel correlations, while encoding an angle as a correlation between edges allows to reuse the same edge as part of many different angles instead of starting from scratch for every angle. Learning hierarchical features can be done by first learning simple features that capture simple (between-pixels) correlations, and then using these features as input to learn more complex features that capture correlations between simple features. This progressive type of feature learning has been introduced by Hinton *et al.* (2006) and has proved very efficient, beating the record on handwritten digit recognition. By adding a sparsity constraint to a hierarchical feature learning algorithm [27] in a Sparse Encoding Symmetric Machine (SESM), we were able to learn features that *individually* capture most of the higher order correlations in an image of a given type, making each feature easier to interpret intuitively than in a distributed coding framework. In Figure. 5, a 2-layer SESM trained on the MNIST dataset of handwritten digits in a totally unsupervised fashion (i.e. without ever giving any label information to the machine) was able to recover recognizable prototypes of the 10 digits: the image shown is the reconstruction of the 10 second-level features learned by the machine. The machine was trained to encode handwritten digit images into 10 units, and decode the 10-unit code to get an image reconstruction as similar as possible to the initial image.

6.3.2. Learning discriminative dictionaries for local image analysis (J. Mairal, F. Bach, J. Ponce, A. Zisserman, joint work with G. Sapiro, University of Minnesota)

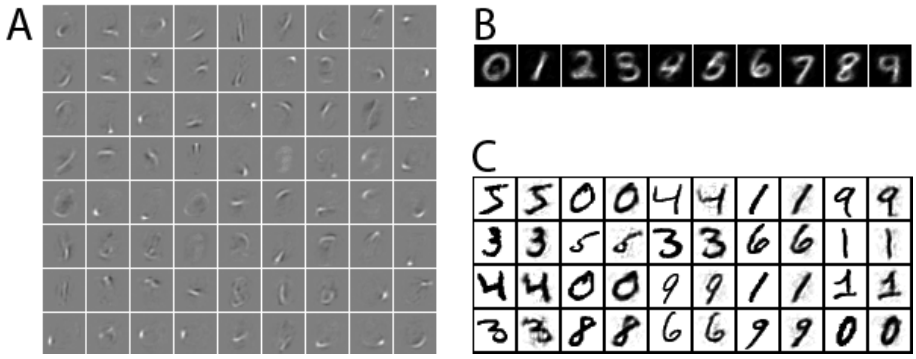

Figure 5. ***(A)*** *Random selection from the 200 linear filters that were learned by the first layer of the SESM* ***(B)*** *Back-projection in image space of the filters learned in the second stage of the hierarchical feature extractor. The second stage was trained on the rectified codes produced by the first stage machine. The back-projection has been performed by using a 1-of-10 code in the second stage machine, and propagating this through the second stage decoder and first stage decoder. The filters at the second stage discover the class-prototypes (manually ordered for visual convenience) even though no class label was ever used during training.* ***(C)*** *Some typical pairs of original and reconstructed digit from the code produced by the encoder in the SESM (feed-forward propagation through encoder and decoder).*

Sparse signal models have been the focus of much recent research, leading to (or improving upon) state-of-the-art results in signal, image, and video *restoration*. We extend this line of research into a novel framework for local image *discrimination* tasks, proposing an energy formulation with both sparse reconstruction and class discrimination components, jointly optimized during dictionary learning. This approach improves over the state of the art in texture segmentation experiments using the Brodatz database, and it paves the way for a novel scene analysis and recognition framework based on simultaneously learning discriminative and reconstructive dictionaries. Preliminary results in this direction using examples from the Pascal VOC06 and Graz02 datasets are promising (Figure 6).

6.3.3. Learning visual attributes (A. Zisserman, joint work with V. Ferrari, Oxford)

We have also recently proposed a probabilistic generative model of visual attributes, together with an efficient learning algorithm. Attributes are visual qualities of objects, such as ‘red’, ‘striped’, or ‘spotted’. The model sees attributes as patterns of image segments, repeatedly sharing some characteristic properties. These can be any combination of appearance, shape, or the layout of segments within the pattern. Moreover, attributes with general appearance are taken into account, such as the pattern of alternation of any two colors which is characteristic for stripes. To enable learning from unsegmented training images, the model is learnt discriminatively, by optimizing a likelihood ratio. As demonstrated by our experimental evaluation, our model can learn in a weakly supervised setting and encompasses a broad range of attributes. We show that attributes can be learnt starting from a text query to Google image search, and can then be used to recognize the attribute and determine its spatial extent in novel real-world images.

6.3.4. Flexible object models for category-level 3D object recognition (A. Kushal and J. Ponce, joint work with C. Schmid, LEAR)

After image and object attribute models, we propose here to learn a part-based object model from visual data. Today’s category-level object recognition systems largely focus on centered fronto-parallel views of nearly rigid objects with characteristic texture patterns. To overcome these limitations, we have proposed in [25] a novel framework for visual object recognition where object classes are represented by assemblies of *partial*

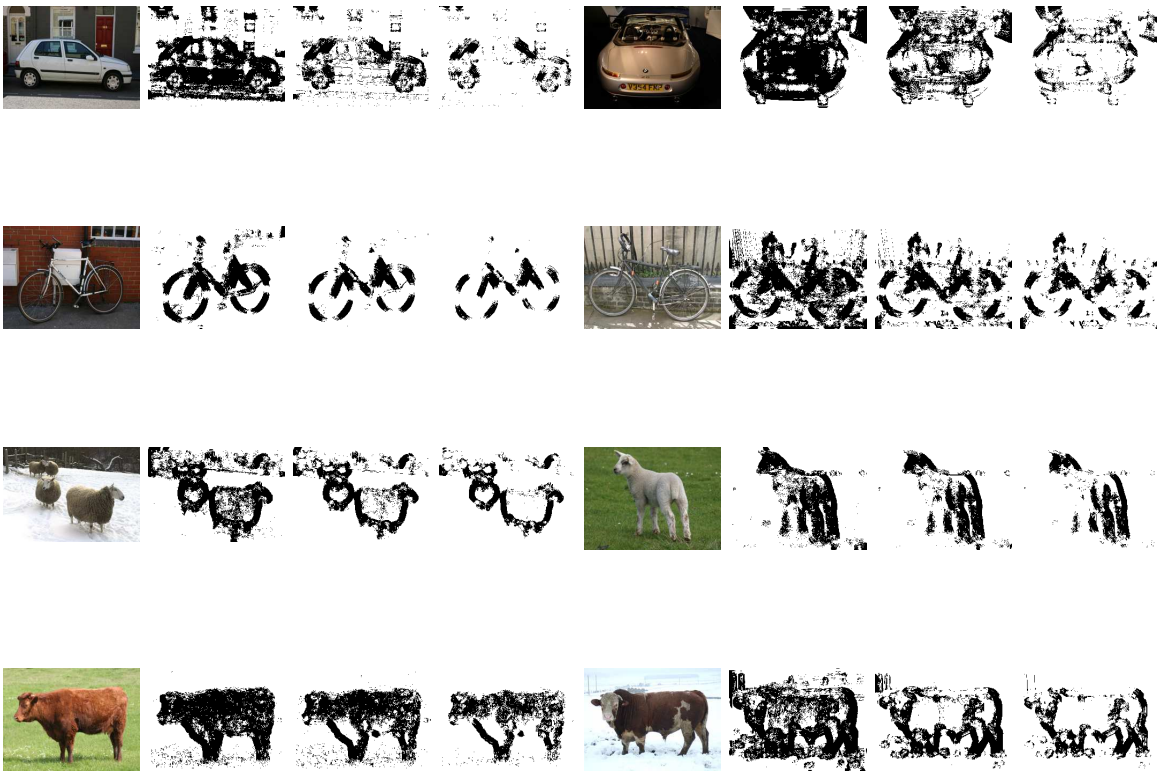

Figure 6. Discriminative feature selected after different numbers of iterations of our algorithm on images from the Pascal'06 dataset.

surface models (PSMs) obeying loose local geometric constraints (Figure 7). The PSMs are formed of dense, locally rigid assemblies of image features. Since our model only enforces *local* geometric consistency, both at the level of model parts and at the level of individual features within the parts, it is robust to viewpoint changes and intra-class variability. The proposed approach has been implemented, and it outperforms the state-of-the-art algorithms for localization recently compared in Everingham et al. (2005) using the Pascal 2005 VOC Challenge Cars Test 1 data.

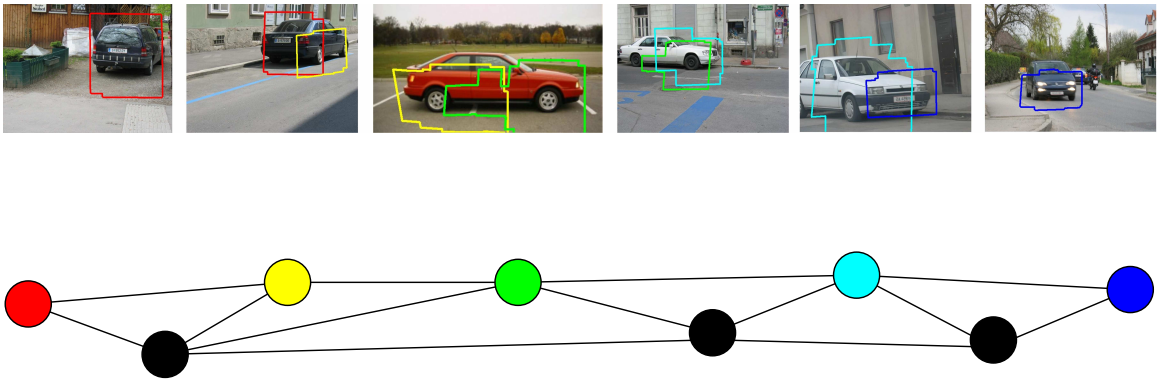

Figure 7. An example of learned PSM graph. The top row shows the outlines of the PSM instances corresponding to nodes with the same color in the PSM graph below it. The black nodes represent other nodes in the PSM graph.

6.3.5. *Unsupervised Discovery of Visual Object Class Hierarchies (J. Sivic, B. Russell, A. Zisserman, joint work with A. Efros (CMU, équipe associée) and B. Freeman (MIT))*

Objects in the world can be arranged into a hierarchy based on their semantic meaning (e.g. organism – animal – feline – cat). But what about defining a hierarchy based on the visual appearance of objects? This paper investigates ways to automatically discover a hierarchical structure for the visual world from a collection of unlabeled images. Previous approaches for unsupervised object and scene discovery focused on partitioning the visual data into a set of nonoverlapping classes of equal granularity. In this work, we propose to group visual objects using a multi-layer hierarchy tree that is based on common visual elements. This is achieved by adapting to the visual domain the generative Hierarchical Latent Dirichlet Allocation (hLDA) model previously used for unsupervised discovery of topic hierarchies in text. Images are modelled using quantized local image regions as analogues to words in text. We demonstrate that meaningful object hierarchies can be automatically learned from unlabelled image collections without supervision. The quality of the hierarchy is assessed in two ways: first, by measuring its classification performance, and second, by automatically discovering objects and their segmentation from an unsegmented and unlabelled set of images. Results are compared with the method of Russell *et al.* CVPR 2006.

6.4. Image Segmentation

6.4.1. *Segmentation with shape priors (P. Etyngier and F. Ségonne, joint work with R. Keriven, ENPC)*

In this work, we introduce a non-linear shape prior for the deformable model framework that can be acquired from shape samples using recent manifold learning techniques [20], [19]. We model a category of shapes as a finite dimensional manifold which we approximate using Diffusion maps, that we call the shape prior manifold. Our method computes a Delaunay triangulation of the reduced space, considered as Euclidean, and

uses the resulting space partition to identify the closest neighbors of any given shape based on its Nyström extension. Our contribution lies in three aspects. First, we propose a solution to the pre-image problem and define the projection of a shape onto the manifold. Based on closest neighbors for the diffusion distance, we then describe a variational framework for manifold denoising. Finally, we introduce a shape prior term for the deformable framework through a non-linear energy term designed to attract a shape towards the manifold at given constant embedding. Segmentation results on shapes of cars and ventricule nuclei are presented and demonstrate the potentials of our method (Fig. 8).

Figure 8. Segmentation of a Peugeot 206 (left) and a Suzuki Swift (right) with data term only (first images) and with shape prior (second images).

6.4.2. Learning to improve a local scene segmentation through global features (J. Ponce, joint work with K. McHenry, UIUC, and S. Lazebnik, UNC)

Local information cannot capture all of the image/scene constraints available for image segmentation. Imposing global constraints such as shape priors as described above can often improve segmentation results. Rather than imposing predetermined global constraints, we propose to attempt to learn one. Given images that are initially over-segmented into regions of nearly uniform color and texture we use a set of global features on these regions and their class assignments to learn an energy function. This energy-based model is trained so as to assign lower energies to segmentations that have a larger percentage of correctly labeled pixels. The resulting energy function is then used to refine a given segmentation constructed from local features of the initial (over-segmented) regions. This approach is simpler than a MRF approach as it uses a single global feature vector instead of a complex energy function composed of a variable number of interaction potentials. We demonstrate our approach with quantitative and qualitative results.

6.4.3. Some links between min-cuts, optimal spanning forests and watersheds (J.-Y. Audibert, joint work with C. Allène, M. Couprie, J. Cousty and R. Keriven)

Different optimal structures (e.g., minimum cuts, minimum spanning forests and shortest-path forests) have been used as the basis for powerful image segmentation procedures. The well-known notion of watershed also falls into this category. In [15], we present some new results about the links which exist between these different approaches. In particular, we show that min-cuts coincide with watersheds for some particular weight functions.

6.5. Machine learning for computer vision

6.5.1. Interactive segmentation by transduction (J.-Y. Audibert, F. Ségonne, J. Ponce, joint work with R. Keriven and O. Duchenne)

Interactive segmentation is a computer vision problem where machine learning offers new insights. Concretely, we address the problem of segmenting an image into regions consistent with user-supplied seeds (e.g., a sparse set of broad brush strokes). We view this task as a statistical *transductive inference*, in which some pixels are already associated with given zones and the remaining ones need to be classified. Our method relies on the Laplacian graph regularizer, a powerful manifold learning tool that is based on the estimation of variants of the Laplace-Beltrami operator and is tightly related to diffusion processes. Segmentation is modeled as the task

of finding matting coefficients for unclassified pixels given known matting coefficients for seed pixels. The proposed algorithm essentially relies on a high margin assumption in the space of pixel characteristics. It is simple, fast, and accurate, as demonstrated by qualitative results on natural images (Figure 9) and a quantitative comparison with state-of-the-art methods on the Microsoft GrabCut segmentation database.

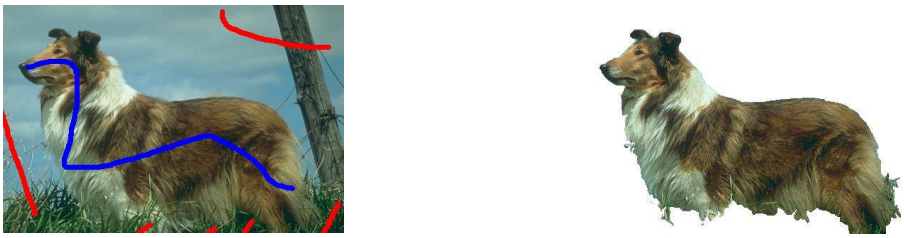

Figure 9. Left: an input image with user-supplied strokes. Right: the segmentation found by our algorithm.

6.5.2. Graph-based methods for interactive image search (J.-Y. Audibert, joint works with H. Sahbi, P. Etyngier and R. Keriven)

Closing the semantic gap in content based image retrieval basically requires the knowledge of the user's intention which is usually translated into a sequence of questions and answers. The user's feedback to these questions provides a partial labeling of the data and makes it possible to iteratively refine a decision rule on the unlabeled data. Training of this decision rule is referred to as transductive learning. In [28], we propose an original approach to relevance feedback based on graph-cuts. Training consists in implicitly modeling the manifold enclosing both the labeled and unlabeled dataset and finding a partition of this manifold using a min-cut. This relevance feedback model exploits the structure of the manifold by considering also the structure of the unlabeled data. Experiments conducted on generic as well as specific databases show that our graph-cut based approach is very effective, outperforms other existing methods and makes it possible to converge to almost all the images of the user's "class of interest" with a very small labeling effort. In [32], we consider the graph Laplacian operator to perform the interactive image search. We introduce a new Graph Laplacian which makes it possible to robustly learn the embedding, of the manifold enclosing the dataset, via a diffusion map. Our approach is three-folds : it allows us (i) to integrate all the unlabeled images in the decision process (ii) to robustly capture the topology of the image set and (iii) to perform the search process inside the manifold. Relevance feedback experiments were conducted on simple databases including Olivetti and Swedish as well as challenging and large scale databases including Corel. Comparisons show clear and consistent gain, of our graph Laplacian method, with respect to state-of-the art relevance feedback approaches.

6.6. Effective learning algorithms and architectures

6.6.1. DIFFRAC : a discriminative and flexible framework for clustering (F. Bach, joint work with Z. Harchaoui, Telecom Paris)

Many clustering frameworks have already been proposed, with numerous applications in machine learning, exploratory data analysis, computer vision and speech processing. However, these unsupervised learning techniques have not reached the level of sophistication of supervised learning techniques, that is, for all methods, there are still a significant number of explicit or implicit parameters to tune for successful clustering, most generally, the number of clusters and the metric or the similarity structure over the space of configurations. In this work, we present a discriminative and flexible framework for clustering (DIFFRAC), which is aimed at alleviating some of those practical annoyances. More precisely, we developed a novel linear clustering framework which relies on a linear discriminative cost function and a convex relaxation of a combinatorial optimization

problem. The large convex optimization problem is solved through a sequence of lower dimensional singular value decompositions. This framework has several attractive properties: (1) although apparently similar to K-means, it exhibits superior clustering performance than K-means, in particular in terms of robustness to noise. (2) It can be readily extended to non linear clustering if the discriminative cost function is based on positive definite kernels, and can then be seen as an alternative to spectral clustering. (3) Prior information on the partition is easily incorporated, leading to state-of-the-art performance for semi-supervised learning, for clustering or classification. We have undertaken empirical evaluations of our algorithms on synthetic and real medium-scale datasets.

6.6.2. Testing for homogeneity with kernel Fisher discriminant analysis (F. Bach, joint work with E. Moulines and Z. Harchaoui, Telecom Paris)

An important problem in statistics and machine learning consists in testing whether the distributions of two random variables are identical under the alternative that they may differ in some ways. This problem arises in many applications, ranging from computational anatomy to process monitoring. We propose to investigate test statistics for testing homogeneity in reproducing kernel Hilbert spaces. Asymptotic null distributions under null hypothesis are derived, and consistency under fixed and local alternatives is assessed. Finally, experimental evidence of the performance of the proposed approach on both artificial and real datasets is studied.

6.6.3. Optimal solutions for sparse principal component analysis (F. Bach, joint work with A. d'Aspremont, Princeton University, and L. El Ghaoui, UC Berkeley)

Principal component analysis (PCA) is a classic tool for data analysis, visualization or compression and has a wide range of applications throughout science and engineering. One of the key shortcomings of PCA is that the factors are linear combinations of all original variables; that is, most of factor coefficients (or loadings) are non-zero. This means that while PCA facilitates model interpretation and visualization by concentrating the information in a few factors, the factors themselves are still constructed using all variables, hence are often hard to interpret. Solutions that have only a few non-zero coefficients in the principal components are usually easier to interpret. Given a sample covariance matrix, we thus examine the problem of maximizing the variance explained by a linear combination of the input variables while constraining the number of nonzero coefficients in this combination. This is known as sparse principal component analysis and has a wide array of applications in machine learning and engineering. We formulate a new semidefinite relaxation to this problem and derive a greedy algorithm that computes a full set of good solutions for all target numbers of non zero coefficients, with total complexity $O(n^3)$, where n is the number of variables. We then use the same relaxation to derive sufficient conditions for global optimality of a solution, which can be tested in $O(n^3)$ per pattern. We discuss applications in subset selection and sparse recovery and show on artificial examples that our algorithm does provide globally optimal solutions in many cases.

6.6.4. Exploration-exploitation trade-off (J.-Y. Audibert, joint work with R. Munos and C. Szepesvari)

Algorithms based on upper-confidence bounds for balancing exploration and exploitation are gaining popularity since they are easy to implement, efficient and effective. In [17], [30], we consider a variant of the basic algorithm for the stochastic, multi-armed bandit problem that takes into account the empirical variance of the different arms. In earlier experimental works, such algorithms were found to outperform the competing algorithms. The purpose of this work is to provide a theoretical explanation of these findings and provide theoretical guidelines for the tuning of the parameters of these algorithms. For this we analyze the expected regret and for the first time the concentration of the regret. The analysis of the expected regret shows that variance estimates can be especially advantageous when the payoffs of suboptimal arms have low variance. The risk analysis, rather unexpectedly, reveals that except for some very special bandit problems, the regret, for upper confidence bounds based algorithms with standard bias sequences, concentrates only at a polynomial rate. Hence, although these algorithms achieve logarithmic expected regret rates, they seem less attractive when the risk of suffering much worse than logarithmic regret is also taken into account.

6.7. Learning theory

6.7.1. Convergence of graph Laplacians (J.-Y. Audibert, joint work with M. Hein and U. von Luxburg)

Given a sample from a probability measure with support on a submanifold in Euclidean space one can construct a neighborhood graph which can be seen as an approximation of the submanifold. The graph Laplacian of such a graph is used in several machine learning methods like semi-supervised learning, dimensionality reduction and clustering. In [8], we determine the pointwise limit of three different graph Laplacians used in the literature as the sample size increases and the neighborhood size approaches zero. We show that for a uniform measure on the submanifold all graph Laplacians have the same limit up to constants. However in the case of a non-uniform measure on the submanifold only the so called random walk graph Laplacian converges to the weighted Laplace-Beltrami operator.

6.7.2. Performing classification by plugging regression estimates (J.-Y. Audibert, joint work with A. Tsybakov)

It has been recently shown that, under the margin (or low noise) assumption, there exist classifiers attaining fast rates of convergence of the excess Bayes risk, i.e., the rates faster than $n^{-1/2}$. The works on this subject suggested the following two conjectures: (i) the best achievable fast rate is of the order n^{-1}, and (ii) the plug-in classifiers generally converge slower than the classifiers based on empirical risk minimization. In [3], we show that both conjectures are not correct. In particular, we construct plug-in classifiers that can achieve not only the fast, but also the *super-fast* rates, i.e., the rates faster than n^{-1}. We establish minimax lower bounds showing that the obtained rates cannot be improved.

6.7.3. Predicting as well as the best expert (J.-Y. Audibert)

In [16], we consider the learning task consisting in predicting as well as the best function in a finite reference set G up to the smallest possible additive term. If $R(g)$ denotes the generalization error of a prediction function g, under reasonable assumptions on the loss function (typically satisfied by the least square loss when the output is bounded), it is known that the progressive mixture rule g_n satisfies $ER(g_n) \leq \min_{g \in G} R(g) + Cst \, \frac{\log |G|}{n}$, where n denotes the size of the training set, and E denotes the expectation with respect to the training set distribution.denotes a positive constant. This work shows that, surprisingly, for appropriate reference sets G, the deviation convergence rate of the progressive mixture rule is no better than $Cst \, / \sqrt{n}$: it fails to achieve the expected $Cst \, / n$. We also provide an algorithm which does not suffer from this drawback, and which is optimal in both deviation and expectation convergence rates.

6.7.4. Consistency of trace norm minimization (F. Bach)

In recent years, regularization by various non Euclidean norms has seen considerable interest. In [5], we consider the rank consistency of trace norm regularization with the square loss, i.e., if the data were actually generated by a low-rank matrix, will the matrix and its rank be consistently estimated? More precisely, we extend some of the consistency results of the Lasso to provide necessary and sufficient conditions for rank consistency of trace norm minimization with the square loss. We also provide an adaptive version that is rank consistent even when the necessary condition for the non adaptive version is not fulfilled.

6.7.5. Consistency of the group Lasso and multiple kernel learning (F. Bach)

Regularization has emerged as a dominant theme in machine learning and statistics. It provides an intuitive and principled tool for learning from high-dimensional data. In recent years, regularization by non Hilbertian norms has generated considerable interest in linear supervised learning, where the goal is to predict a response as a linear function of covariates; in particular, regularization by the L1-norm (the sum of absolute values), a method commonly referred to as the Lasso (Tibshirani, 1994, Osborne et al., 2000), allows to perform variable selection. In [4], we extend the consistency results of the Lasso to the group Lasso, by studying the asymptotic model consistency of the group Lasso. We derive necessary and sufficient conditions for the consistency of group Lasso under practical assumptions, such as model misspecification. When the linear predictors and

Euclidean norms are replaced by functions and reproducing kernel Hilbert norms, the problem is usually referred to as multiple kernel learning and is commonly used for learning from heterogeneous data sources and for non linear variable selection. Using tools from functional analysis, and in particular covariance operators, we extend the consistency results to this infinite dimensional case and also propose an adaptive scheme to obtain a consistent model estimate, even when the necessary condition required for the non adaptive scheme is not satisfied.

7. Contracts and Grants with Industry

7.1. Introduction

Since the members of WILLOW belong to different institutions, some of our grants are managed by INRIA, while other are managed by ENS or ENPC. We indicate below the managing institution for each grant.

7.2. DGA/Bertin/EADS/SAGEM: 2ACI (ENS, pending)

Participant: Jean Ponce.

This project is concerned with target detection in low-resolution infra-red images. WILLOW is part of three consortiums involving different industrials (namely, Bertin, EADS, and Sagem) and academic partners (including INRIA). These three consortiums are the three finalists chosen by DGA so there is a high likelihood the WILLOW part, which is concerned with the detection of 3D targets and the estimation of their pose, will be funded. Total WILLOW budget: 110 KEuros.

7.3. DGA/E-vitech: ITISECURE (ENS)

Participants: Jean-Yves Audibert, Jean Ponce.

This contract belongs to our automatic scene understanding research program. It aims at designing unexpected object detection algorithms in the framework of a vehicle moving several times on the same route. The core problems involved by this task are image matching handling high variations in the video capturing conditions and scene understanding (objects identification, position and movement). Several parts of computer vision and machine learning are thus involved: optical flow estimation, image processing, feature extraction and matching in low-dimensional images, hypothesis testing, statistical learning, etc. J.-Y. Audibert is its coordinator. Total WILLOW funding: 60 KEuros.

7.4. EADS (ENS)

Participants: Jean Ponce, Andrew Zisserman.

A. Zisserman's participation in WILLOW has been partially funded through an EADS industrial chair at ENS. This has resulted in initial collaboration efforts via discussions and tutorial presentations by A. Zisserman and J. Ponce at EADS. The tutorial was delivered at EADS Suresnes lab in May 2007. It covered Multiple View Geometry, and in particular the following areas: reconstruction and estimation, projective reconstruction from multiple views, estimation of the fundamental matrix and calibration.

7.5. MSR-INRIA joint lab: Image and video mining for science and humanities (INRIA)

Participants: Jean Ponce, Francis Bach, Andrew Zisserman.

This new collaborative project, already mentioned several times in this report, brings together the WILLOW, LEAR, and VISTA project-teams with MSR researchers in Cambridge and elsewhere. The concept builds on several ideas articulated in the "2020 Science" report, including the importance of data mining and machine learning in computational science. Rather than focusing only on natural sciences, however, we propose here to expand the breadth of e-science to include humanities and social sciences. The project we propose will focus on fundamental computer science research in computer vision and machine learning, and its application to archaeology, cultural heritage preservation, environmental science, and sociology, and it will be validated by collaborations with researchers and practitioners in these fields. Total budget: 628 KEuros.

8. Other Grants and Activities

8.1. Agence Nationale de la Recherche: HFIMBR (INRIA)

Participants: Florent Ségonne, Jean Ponce, Jean-Philippe Pons, Andrew Zisserman.

This is a collaborative effort with A. Bartoli (LASMEA Clermont-Ferrand) and N. Holszuch (ARTIS project-team, INRIA Rhône-Alpes).

There is an increasing need for three-dimensional (3D) "content" in entertainment, engineering, and scientific applications. We predict that, for most of these, today's specialized 3D sensors will eventually be replaced by ordinary, consumer-grade digital cameras equipped with advanced image-based modeling and analysis software. We propose core computer vision and computer graphics research that will enable the development of this software and its application to real-world problems. Concretely, we will focus on high-fidelity image-based modeling and 3D shape and appearance matching, and we will demonstrate applications of the technology developed in this project to film post production and special effects, and cultural heritage conservation, both pursued via collaborations with external partners. Total funding for WILLOW: 110 KEuros.

8.2. Agence Nationale de la Recherche: MGA (INRIA/ENPC)

Participants: Francis Bach, Jean-Yves Audibert, Jean Ponce, Andrew Zisserman.

Probabilistic graphical models, also known as Bayesian Networks, provide a very flexible and powerful framework for capturing statistical dependencies in complex, multivariate data. They enable the building of large global probabilistic models for complex phenomena out of smaller and more tractable local models. The objectives of this project are to advance the methodological state of the art of probabilistic modeling research, while applying the newly developed techniques to computer vision, text processing and bio-informatics. F. Bach is the coordinator of this ANR "projet blanc" in machine learning, that focuses on graphical models and their applications. The total funding is 200 KEuros, with 100KEuros for Willow including (50KEuros for INRIA and 50KEuros for ENPC). The kick-off meeting took place on December 13th, 2007.

8.3. Agence Nationale de la Recherche: Triangles (ENS)

Participant: Jean Ponce.

This is a collaborative effort with O. Devillers (INRIA project-team GEOMETRICA), Raphaelle Chaine (University of Lyon), and J. Ponce and E. Colin de Verdière (ENS).

This project is dedicated to the design of computational geometry methods for constructing triangulation in non-Euclidean spaces. Total funding for WILLOW: 5000 Euros.

8.4. Getty Conservation Research Institute (ENS)

Participants: Jean Ponce, Mariano Tepper.

This project is concerned with the development of software and methodology for assessing the deterioration over time of the stones making up the Maya Hyeroglyphic Stairway in Copan, Honduras. The proposed method uses modern structure-from-motion, registration stereo techniques to compare pairs of images of the stones taken in 2000 and 2004, and assess the damage. Total funding: $5000.

8.5. France-UC Berkeley fund (Ecole des Mines de Paris)

Participant: Francis Bach.

This ia a travel Grant from the French Berkeley fund (http://ies.berkeley.edu/fbf/), joint with Jean-Philippe Vert (Ecole des Mines de Paris) and Michael Jordan (UC Berkeley). Total funding: 10,000 Euros.

9. Dissemination

9.1. Leadership within the scientific community

- Conference and workshop organization:
  - General chair, European Conference on Computer Vision, Marseille, 2008 (J. Ponce).
  - Program chair, European Conference on Computer Vision, Marseille, 2008 (A. Zisserman).
  - Chair, Pascal VOC Challenge Workshop, Rio de Janeiro, 2007 (A. Zisserman).
- Editorial boards:
  - International Journal of Computer Vision (J. Ponce, editor in chief).
  - International Journal of Computer Vision (A. Zisserman).
  - Foundations and Trends in Computer Graphics and Vision (J. Ponce).
- Area chairs:
  - Asian Conference on Computer Vision, 2007 (J. Ponce).
  - International Conference on Computer Vision, 2007 (J. Ponce).
  - Neural Information and Processing Systems (NIPS) Conference, 2007 (F. Bach and A. Zisserman).
- Program committees:
  - IEEE Conference on Computer Vision and Pattern Recognition, 2007 (J. Ponce).
- Other:
  - J.-Y. Audibert is associate member of the PASCAL European Network of Excellence (http://www.pascal-network.org).
  - F. Bach is a member of the PASCAL European Network of Excellence (http://www.pascal-network.org).
  - F. Bach coordinates the ParisTech reading group in machine learning (http://www.di.ens.fr/~fbach/paristech/).
  - J. Koenderink gave four public lectures at ENS (http://www.di.ens.fr/willow/invitedSpeakers.html.
  - J. Ponce is responsible for teaching and the entrance exam in the department of computer science of Ecole normale supérieure.
  - J. Ponce is a member of the scientific advisory board for the Institut de l'Ecole normale supérieure.

- J. Ponce organizes the ENS computer vision seminar (see http://www.di.ens.fr/~ponce/semspring07.html and http://www.di.ens.fr/~ponce/semfall07.html).
- J. Ponce served on the 2007 admission committee for research directors at INRIA.
- J. Ponce and A. Zisserman, in collaboration with Y. Furukawa (UIUC) are starting an effort aimed at reconstructing vases from the Beazley Collection (http://www.beazley.ox.ac.uk/Pottery/Ashmolean/Script/default.htm.
- A. Zisserman is a member of the PASCAL European Network of Excellence and co-organizes the Pascal VOC challenge (http://www.pascal-network.org/challenges/VOC/voc2005/).

9.2. Teaching

- J.-Y. Audibert, “Statistics”, Ecole Nationale des Ponts et Chaussées, 2nd year, 26h.
- J.-Y. Audibert, “Machine Learning”, Masters (M2) “Mathématiques, Vision et Apprentissage” (MVA), Ecole Normale Supérieure de Cachan, 20h.
- F. Bach, “Probabilistic graphical models”, MVA, Ecole Normale Supérieure de Cachan, 20h.
- J. Ponce, “Geometry and computer vision”, Ecole normale supérieure and MVA, Ecole normale supérieure de Cachan, 24h.
- J. Ponce, “Introduction to scientific computing”, Ecole normale supérieure, M1, 36h.
- J.-P. Pons, “Mathematics and Computer Science”, Ecole Nationale des Ponts et Chaussées, 2nd year, 21 h.
- F. Ségonne, “Algorithms and Programming”, Ecole Nationale des Ponts et Chaussées, 2nd year, 63 h.
- F. Ségonne, “Applied Maths and Computer Vision”, Ecole Nationale des Ponts et Chaussées, 2nd year, 28 h.
- A. Zisserman, Third year lecture course on "Estimation and Inference", Oxford.
- A. Zisserman, Third year labs on "Information Engineering", Oxford.
- A. Zisserman, Fourth year lecture course on "Optimization", Oxford.

9.3. Invited presentations

- J.-Y. Audibert, *Fast learning rates for plug-in classifiers*, Empirical Processes and Asymptotic Statistics, Univ. Rennes 1, Jun. 2007
- J.-Y. Audibert, *Convergence of the graph Laplacian: application to dimensionality estimation and image segmentation*, Pascal Workshop on Graph Theory and Machine Learning, Bled, Slovenia, Jun. 2007
- J.-Y. Audibert, *Aggregation to compete the best prediction function in a finite set*, Probability and Statistics in Science and Technology, ISI, Porto, Portugal, Sept. 2007
- J.-Y. Audibert, *Graph-based methods for manifold learning*, Mathematics for biological networks, Institut Henri Poincaré, Paris, Dec. 2007
- J. Ponce, *High-fidelity image- and video-based modeling*, ACCV’07 Workshop, Hiroshima.
- J. Ponce, *High-fidelity image- and video-based modeling*, Microsoft Research, Cambridge.
- J. Ponce, *The challenge of 3D computer vision*, ICCV’07 3D Workshop, Rio de Janeiro.
- J. Ponce and A. Zisserman, *A tutorial on shape from motion and auto-calibration*, EADS.
- J.-P. Pons, *Shape reconstruction from images using some recent Delaunay-based algorithms*, INRIA Sophia-Antipolis, January 2007.

- J.-P. Pons, *Multi-view 3D reconstruction with deformable models and Delaunay meshing*, Ecole Centrale Paris, February 2007.
- J.-P. Pons, *The deformable models framework: Shape reconstruction using moving interfaces in computer vision and image processing*, SciCADE 2007 conference, Saint-Malo, July 2007.
- J.-P. Pons, *Shape recontruction from images using Delaunay meshing: Some recent results*, INRIA Sophia-Antipolis, October 2007.
- J.-P. Pons, *Shape recontruction from images using Delaunay meshing: Some recent results*, ESIEE, Marne-la-Vallée, December 2007.
- F. Ségonne, *Segmentation and topological constraints*, ESIEE, Marne-la-Vallée, June 2007.
- A. Zisserman, Plenary speaker at the Asian Conference on Computer Vision (ACCV) 2006, India
- A. Zisserman, Microsoft Bangalore Computer Vision Workshop 2006
- A. Zisserman, Key note speaker at the ACM International Conference on Image and Video Retrieval, (CIVR) 2007, Amsterdam. http://www.civr2007.com/
- A. Zisserman, Tutorial at the Second Summer School on Multimedia Semantics, Glasgow
- A. Zisserman, IPAM workshop on "Numerical Tools and Fast Algorithms for Massive Data Mining, Search Engines and Applications", UCLA.

10. Bibliography

Year Publications

Books and Monographs

[1] J. PONCE, M. HEBERT, C. SCHMID, A. ZISSERMAN. *Toward Category-Level Object Recognition*, vol. 4170, Springer-Verlag Lecture Notes in Computer Science, 2007.

Articles in refereed journals and book chapters

[2] J.-Y. AUDIBERT, O. BOUSQUET. *Combining PAC-Bayesian and generic chaining bounds*, in "Journal of Machine Learning Research", vol. 8, 2007, p. 863–889.

[3] J.-Y. AUDIBERT, A. TSYBAKOV. *Fast learning rates for plug-in classifiers*, in "Annals of Statistics", vol. 35, n^{o} 2, 2007, p. 608-633.

[4] F. BACH. *Consistency of the group Lasso and multiple kernel learning*, in "Journal of Machine Learning Research", Accepted for publication, 2007.

[5] F. BACH. *Consistency of trace norm minimization*, in "Journal of Machine Learning Research", Submitted, 2007.

[6] J. ERICKSON, S. THITE, F. ROTHGANGER, J. PONCE. *Capturing a Convex Object with Three Discs*, in "IEEE Trans. Robotics", Accepted for publication., 2007.

[7] Y. FURUKAWA, J. PONCE. *Carved Visual Hulls for Image-Based Modeling*, in "Int. J. of Comp. Vision", Submitted., 2007.

[8] M. HEIN, J.-Y. AUDIBERT, U. VON LUXBURG. *Graph laplacians and their convergence on random neighborhood graphs*, in "Journal of Machine Learning Research", vol. 8, 2007, p. 1325–1368.

[9] S. LAZEBNIK, Y. FURUKAWA, J. PONCE. *Projective Visual Hulls*, in "Int. J. of Comp. Vision", vol. 74, n^{o} 2, 2007, p. 137–166.

[10] J. MAIRAL, G. SAPIRO, M. ELAD. *Learning multiscale sparse representations for image and video restoration*, in "SIAM Multiscale Modeling and Simulation", Accepted for publication., 2007.

[11] F. ROTHGANGER, S. LAZEBNIK, C. SCHMID, J. PONCE. *Segmenting, Modeling, and Matching Video Clips Containing Multiple oving Objects*, in "IEEE Trans. Patt. Anal. Mach. Intell.", vol. 29, n^{o} 3, 2007, p. 477-491.

[12] F. SÉGONNE. *Active Contours Under Topology Control - Genus Preserving Level Sets*, in "Int. J. of Comp. Vision", To appear., 2007.

[13] A. D'ASPREMONT, F. BACH, L. E. GHAOUI. *Optimal solutions for sparse principal component analysis*, in "Journal of Machine Learning Research", Submitted, 2007.

Publications in Conferences and Workshops

[14] E. AGANJ, J.-P. PONS, F. SÉGONNE, R. KERIVEN. *Spatio-temporal shape from silhouette using four-dimensional Delaunay meshing*, in "Proc. Int. Conf. Comp. Vision, Rio de Janeiro, Brazil", Oct 2007, http://www.enpc.fr/certis/publications/papers/07iccv_b.pdf.

[15] C. ALLÈNE, J.-Y. AUDIBERT, J. COUPRIE, R. KERIVEN. *Some links between min-cuts, optimal spanning forests and watersheds*, in "Proc. 8th International Symposium on Mathematical Morphology, Rio de Janeiro, Brazil", Oct 2007, http://www.enpc.fr/certis/publications/papers/ISMM07.pdf.

[16] J.-Y. AUDIBERT. *Progressive mixture rules are deviation suboptimal*, in "Advances in Neural Information Processing Systems", vol. 20, Dec 2007.

[17] J.-Y. AUDIBERT, R. MUNOS, C. SZEPESVÁRI. *Tuning bandit algorithms in stochastic environments*, in "18th International Conference on Algorithmic Learning Theory, Japon", Oct 2007.

[18] F. BACH, Z. HARCHAOUI. *DIFFRAC: a discriminative and flexible framework for clustering*, in "Advances in Neural Information Processing Systems (NIPS)", In Press, vol. 20, 2007.

[19] P. ETYNGIER, F. SÉGONNE, R. KERIVEN. *Active-Contour-Based Image Segmentation using Machine Learning Techniques*, in "10th IEEE International Conference on Medical Image Computing and Computer Assisted Intervention, Brisbane, Australia", Oct 2007, p. 891-899, http://www.enpc.fr/certis/publications/papers/MICCAI07.pdf.

[20] P. ETYNGIER, F. SÉGONNE, R. KERIVEN. *Shape priors using Manifold Learning Techniques*, in "Proc. Int. Conf. Comp. Vision, Rio de Janeiro, Brazil", Oct 2007, http://www.enpc.fr/certis/publications/papers/07iccv_c.pdf.

[21] A. FARAHMAND, C. SZEPESVÁRI, J.-Y. AUDIBERT. *Manifold-adaptive dimension estimation*, in "Proceedings of the 24th International conference on Machine Learning, Oregon , USA", Jun 2007.

[22] A. FARAHMAND, C. SZEPESVÁRI, J.-Y. AUDIBERT. *Toward Manifold-Adaptive Learning*, in "NIPS Workshop on Topology learning", Dec 2007.

[23] Y. FURUKAWA, J. PONCE. *Accurate, Dense, and Robust Multi-View Stereopsis*, in "Proc. IEEE Conf. Comp. Vision Patt. Recog.", 2007.

[24] Z. HARCHAOUI, F. BACH, E. MOULINES. *Testing for Homogeneity with Kernel Fisher Discriminant Analysis*, in "Advances in Neural Information Processing Systems (NIPS)", In Press, vol. 20, 2007.

[25] A. KUSHAL, C. SCHMID, J. PONCE. *Flexible Object Models for Category-Level 3D Object Recognition*, in "Proc. IEEE Conf. Comp. Vision Patt. Recog.", 2007.

[26] P. LABATUT, J.-P. PONS, R. KERIVEN. *Efficient multi-view reconstruction of large-scale scenes using interest points, Delaunay triangulation and graph cuts*, in "Proc. Int. Conf. Comp. Vision, Rio de Janeiro, Brazil", Oct 2007, http://www.enpc.fr/certis/publications/papers/07iccv_a.pdf.

[27] M. RANZATO, Y-LAN. BOUREAU, Y. LECUN. *Sparse feature learning for deep belief networks*, in "Advances in Neural Information Processing Systems (NIPS)", 2007.

[28] H. SAHBI, J.-Y. AUDIBERT, R. KERIVEN. *Graph-cut transducers for relevance feedback in content based image retrieval*, in "Proc. Int. Conf. Comp. Vision, Rio de Janeiro, Brazil", Oct 2007, http://www.enpc.fr/certis/publications/papers/07iccv_d.pdf.

Internal Reports

[29] J.-Y. AUDIBERT. *No fast exponential deviation inequalities for the progressive mixture rule*, Technical report, n^{o} 07-35, CERTIS, Mar 2007, http://www.enpc.fr/certis/publications/papers/07certis35.pdf.

[30] J.-Y. AUDIBERT, R. MUNOS, C. SZEPESVÁRI. *Variance estimates and exploration function in multi-armed bandit*, Technical report, n^{o} 07-31, CERTIS, Mar 2007, http://www.enpc.fr/certis/publications/papers/07certis31.pdf.

[31] H. SAHBI, J.-Y. AUDIBERT, R. KERIVEN. *Graph-cut transducers for relevance feedback in content based image retrieval*, Technical report, n^{o} 07-30, CERTIS, Feb 2007, http://www.enpc.fr/certis/publications/papers/07certis30.pdf.

[32] H. SAHBI, P. ETYNGIER, J.-Y. AUDIBERT, R. KERIVEN. *Graph Laplacian for Interactive Image Retrieval*, Technical report, n^{o} 07-32, CERTIS, Apr 2007, http://www.enpc.fr/certis/publications/papers/CERTIS0732.pdf.