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

Activity Report

2008

Table of contents

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

1. Team

Research Scientist

Jean Ponce [Team Leader, 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]
Andrew Zisserman [Team Co-leader, Professor in the Engineering Department of the University of Oxford, and part-time professor at ENS, HdR]
Sylvain Arlot [Chargé de Recherches CNRS]
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]
Josef Sivic [Chargé de Recherches INRIA]

PhD Student

Y-Lan Boureau
Olivier Duchenne
Loic Fevrier
Rodolphe Jenatton
Julien Mairal
Marc Sturzel
Oliver Whyte

Post-Doctoral Fellow

Bryan Russell
Jan van Gemert
Hui Kong

Visiting Scientist

Alexei Efros [Assistant Professor in the Robotics Institute and Computer Science Department of the Carnegie Mellon University]

Administrative Assistant

Nathalie Abiola

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 is a joint research team between INRIA Paris Rocquencourt, Ecole Normale Supérieure (ENS) and Centre National de la Recherche Scientifique (CNRS). This year we have hired two new researchers: Josef Sivic ("chargé de recherche", INRIA) has joined WILLOW in March 2008 and Sylvain Arlot ("chargé de recherche", CNRS) has joined WILLOW in September 2008. In addition, we have hired two post-docs: Jan van Gemert (DGA) and Hui Kong (DGA), and five new PhD students: Olivier Duchenne (ENS), Loic Fevrier (ENS), Rodolphe Jenatton (INRIA), Marc Sturzel (EADS) and Oliver Whyte (INRIA). Three students of Jean Ponce at UIUC have graduated: Y. Furukawa (2008, now a post-doc a the University of Washington), A. Kushal (2008, now with Two Sigma Investments), K. McHenry (2008, just graduated). Alexei Efros (Professor, Carnegie Mellon University, USA) visited WILLOW together with two students for 5 months in Spring 2008.

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, physical and statistical models of materials and illumination patterns, and modeling and retrieval of objects and scenes in large image collections. Our past work in these areas includes research aimed at recognizing rigid 3D objects in cluttered photographs taken from arbitrary viewpoints (Rothganger *et al.*, 2006), segmenting video sequences into parts corresponding to rigid scene components before recognizing these in new video clips (Rothganger *et al.*, 2007), and retrieval of particular objects and buildings from images and videos (Sivic and Zisserman, 2003) and (Philbin *et al.*, 2007). Our current research focuses on acquisition of detailed object models from multiple images and video streams, theoretical analysis of camera models, and object/scene retrieval.

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

We have recently developed several algorithms for multi-view stereopsis (Furukawa and Ponce, 2007) 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. Some of the corresponding software (PMVS,http://www-cvr.ai.uiuc.edu/~yfurukaw/research/pmvs/index.html) is available for free for academics, and licensing negotiations with several companies are under way. We have also recently developed a new calibration algorithm that uses rough multi-view reconstructions to obtain extremely accurate intrinsic and extrinsic camera parameters. Finally, our research has also focused on theoretical analysis of camera models using the formalism and terminology of classical projective geometry.

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 and locally rigid but globally nonrigid models of surface motion to accurately capture the deforming shape of the observed surfaces.

3.1.3. *Retrieval and modeling of objects and scenes in large image collections*

We have recently developed large-scale object retrieval algorithms, which employ soft-assignment of local image descriptors to multiple quantized codewords, thus mitigating some of the quantization effects observed in previous methods and improving on their retrieval performance. We have also introduced a Geometric Latent Dirichlet Allocation (gLDA) model for unsupervised modeling of unordered image collections and investigated scene category recognition techniques for navigating and exploring large collections of still images.

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.

3.2.1. *Learning image and object models.*

We have continued our research on learning sparse image representations (e.g. Mairal *et al., 2007*) by generalizing the previously developed model to discriminative image understanding tasks such as texture segmentation, category-level edge selection and image classification. We have also investigated optimization methods for learning the sparse image representations.

Another significant strand of our research has focused on the extremely challenging task of category-level object/scene matching and alignment. We have developed methods based on (i) graph matching and (ii) discrete optimization of displacement fields. The former method is able to match complex silhouettes of different instances of the same object category, whereas the latter approach can robustly align complicated scenes with large spatial distortions.

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.

3.3. Human activity capture and classification

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.

3.3.1. *Naming and recognition of characters in TV video*

We have recently extended our previous work on automatic naming of characters in videos (Everingham, Sivic, Zisserman, 2006), which considered only frontal faces, by introducing detection, tracking and recognition of characters in profile views, thereby significantly increasing the proportion of video labelled. We have also demonstrated improved recognition performance by learning character-specific classifiers able to automatically learn features discriminating between the different characters present in the video.

3.4. Machine learning

3.4.1. *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 and image retrieval 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.

3.4.2. *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.

3.4.3. *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.

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 (Furukawa and Ponce, 2007) 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.

5.5. Resampling Penalization for histogram selection in regression software

Resampling Penalization is a family of model selection procedure by penalization that can use any exchangeable weighted bootstrap resampling scheme to compute a penalty. It is properly defined in the general framework and extensively studied for histogram selection in regression in [43]. This software is a Matlab package allowing to perform Resampling Penalization for several examples of weights in the histogram selection case. The Resampling Penalization package is provided free for non-commercial use under the terms of the GNU General Public License. It is publicly available at the url http://www.di.ens.fr/~arlot/code/RP.htm.

6. New Results

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

6.1.1. What is a camera? (J.Ponce)

We address in [35] the problem of characterizing a general class of cameras under reasonable, “linear” assumptions (Figure 1). Concretely, we use the formalism and terminology of classical projective geometry to model cameras by two-parameter linear families of straight lines—that is *reguli* (rank-3 families) and *linear congruences* (rank-4 families). This model captures both the *general linear cameras* of Yu and McMillan and the *linear oblique cameras* of Pajdla. From a geometric perspective, it affords a simple classification of all possible camera configurations. From an analytical viewpoint, it also provides a simple and unified methodology for deriving general formulas for projection and inverse projection, triangulation, and binocular and trinocular geometry.

6.1.2. 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 have proposes in [23] 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.

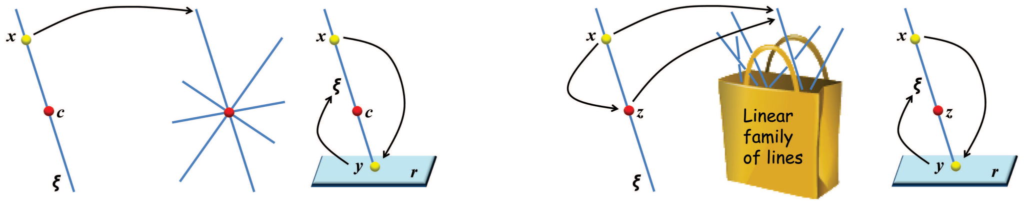

Figure 1. A pinhole camera (left) can be thought of as a device that associates with any point $\mathbf{x}$ *the ray* ξ *that joins it to its image and passes through the pinhole* $\mathbf{c}$*. This ray is picked from the bundle of lines passing through* $\mathbf{c}$*. More generally, a (non-central) camera (right) can be modeled as a device that picks a line from a linear "bag of lines"—that is, a regulus of a linear congruence.*

6.2. Video-based modeling of deformable surfaces

6.2.1. Dense 3D motion capture for human faces (J. Ponce, joint work with Y. Furukawa (Univeristy of Washington)

We propose in [25] a novel approach to motion capture from multiple, synchronized video streams, specifically aimed at recording dense and accurate models of the structure and motion of highly deformable surfaces such as skin, that stretches, shrinks, and shears in the of midst of normal facial expressions. Solving this problem is a key step toward effective performance capture for the entertainment industry, but progress so far has been hampered by the lack of appropriate local motion and smoothness models. The main technical contribution of this paper is a novel approach to regularization adapted to nonrigid tangential deformations. Concretely, we first estimate undergoing nonrigid tangential surface deformation at each vertex of a surface mesh, then aggregate the estimated deformation parameters over the surface for robustness. The estimated deformation parameters are then used in regularizing the (tangential) motion information. To demonstrate the power of the proposed approach, we have integrated it into the state-of-the-art approach to markerless motion capture developed by Furukawa and Ponce [24], and compared the performances of the original and new algorithms on three extremely challenging face datasets that include highly nonrigid skin deformations, wrinkles, and quickly changing expressions. We have also tested the proposed approach on a dataset featuring fast-moving cloth without stretch, shrink or shear, but very complicated and evolving fold structures, and demonstrate the robustness of our new regularization scheme.

6.3. Retrieval and modeling of objects and scenes in large image collections

6.3.1. Improving Particular Object Retrieval in Large Scale Image Databases (J. Sivic and A. Zisserman, joint work with J. Philbin (Oxford), O. Chum (CTU Prague), M. Isard (Microsoft))

The state of the art in visual object retrieval from large databases is achieved by systems that are inspired by text retrieval. A key component of these approaches is that local regions of images are characterized using high-dimensional descriptors which are then mapped to visual words selected from a discrete vocabulary. This work [33] explores techniques to map each visual region to a weighted set of words, allowing the inclusion of features which were lost in the quantization stage of previous systems. The set of visual words is obtained by selecting words based on proximity in descriptor space. We describe how this representation may be incorporated into a standard tf-idf architecture, and how spatial verification is modified in the case of this soft-assignment. We evaluate our method on the standard Oxford Buildings dataset, and introduce a new dataset for evaluation. Our results exceed the current state of the art retrieval performance on these datasets, particularly on queries with poor initial recall where techniques like query expansion suffer. Overall we show

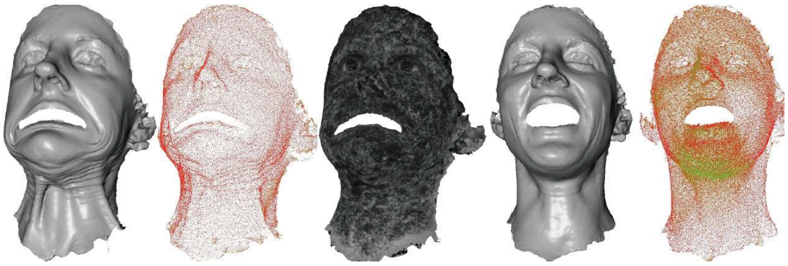

Figure 2. Facial motion capture [25], featuring shaded renderings of reconstructions obtained from two different frames, the corresponding dense motion fields, and one texture-mapped rendering (the actress's face was covered with make-up to provide additional texture). See http://www.cs.washington.edu/homes/furukawa/gallery/ for videos. Data courtesy of Image Movers Digital.

that soft-assignment is always beneficial for retrieval with large vocabularies, at a cost of increased storage requirements for the index.

6.3.2. Geometric LDA: A Generative Model for Particular Object Discovery (J. Sivic and A. Zisserman, joint work with J. Philbin (Oxford))

Automatically organizing collections of images presents serious challenges to the current state-of-the art methods in image data mining. Often, what is required is that images taken in the same place, of the same thing, or of the same person be conceptually grouped together.

To achieve this, we introduce the Geometric Latent Dirichlet Allocation (gLDA) model [34] for unsupervised particular object discovery in unordered image collections. This explicitly represents documents as mixtures of particular objects or facades, and builds rich latent topic models which incorporate the identity and locations of visual words specific to the topic in a geometrically consistent way. Applying standard inference techniques to this model enables images likely to contain the same object to be probabilistically grouped and ranked.

We demonstrate the model on a publicly available dataset of Oxford images, and show examples of spatially consistent groupings.

6.3.3. Creating and Exploring a Large Photorealistic Virtual Space (J. Sivic, joint work with B. Kaneva (MIT), A. Torralba (MIT), S. Avidan (Adobe Research), W.T. Freeman (MIT))

We present a system [38] for exploring large collections of photos in a virtual 3D space. Our system does not assume the photographs are of a single real 3D location, nor that they were taken at the same time. Instead, we organize the photos in themes, such as city streets or skylines, and let users navigate within each theme using intuitive 3D controls that include move left/right, zoom and rotate. Themes allow us to maintain a coherent semantic meaning of the tour, while visual similarity allows us to create a "being there" impression, as if the images were of a particular location. We present results on a collection of several million images downloaded from Flickr and broken into themes that consist of a few hundred thousand images each. A byproduct of our system is the ability to construct extremely long panoramas, as well as image taxi, a program that generates a virtual tour between a user supplied start and finish images. The system, and its underlying technology can be used in a variety of applications such as games, movies and online virtual 3D spaces like Second Life.

6.4. Learning image and object models

6.4.1. Discriminative Sparse Image Models for Class-Specific Edge Detection and Image Interpretation (J. Mairal, F. Bach, M. Hebert and J. Ponce, in collaboration with M. Leordeanu and M. Hebert, Carnegie Mellon University)

Sparse signal models learned from data are widely used in audio, image, and video restoration. We recently generalized them to discriminative image understanding tasks such as texture segmentation and feature selection [29]. This work extends this line of research by proposing a multiscale method to minimize least-squares reconstruction errors and discriminative cost functions under ℓ_0 or ℓ_1 regularization constraints. It is applied to edge detection, category-based edge selection and image classification tasks. Experiments on the Berkeley edge detection benchmark and the PASCAL VOC'05 and VOC'07 datasets demonstrate the computational efficiency of our algorithm and its ability to learn local image descriptions that effectively support demanding computer vision tasks (see figure 3).

6.4.2. Supervised Dictionary Learning (J. Mairal, F. Bach, J. Ponce, A. Zisserman, in collaboration with G. Sapiro, University of Minnesota)

It is now well established that sparse signal models are well suited to restoration tasks and can effectively be learned from audio, image, and video data. Recent research has been aimed at learning discriminative sparse models instead of purely reconstructive ones. Our work proposes a new step in that direction, with a novel sparse representation for signals belonging to different classes in terms of a shared dictionary and multiple discriminative class models. The linear variant of the proposed model admits a simple probabilistic interpretation, while its most general variant admits an interpretation in terms of kernels. An optimization framework for learning all the components of the proposed model is presented, along with experimental results on standard handwritten digit and texture classification tasks. This work extends our previous approach [29], using the coefficients of the sparse decompositions as learned features [30].

6.4.3. A path following algorithm for the graph matching problem (F. Bach, in collaboration with M. Zaslavskiy and J.-P. Vert, Ecole des Mines de Paris)

We propose a convex-concave programming approach for the labeled weighted graph matching problem. The convex-concave programming formulation is obtained by rewriting the weighted graph matching problem as a least-square problem on the set of permutation matrices and relaxing it to two different optimization problems: a quadratic convex and a quadratic concave optimization problem on the set of doubly stochastic matrices. The concave relaxation has the same global minimum as the initial graph matching problem, but the search for its global minimum is also a hard combinatorial problem. We therefore construct an approximation of the concave problem solution by following a solution path of a convex-concave problem obtained by linear interpolation of the convex and concave formulations, starting from the convex relaxation. This method allows to easily integrate the information on graph label similarities into the optimization problem, and therefore to perform labeled weighted graph matching. The algorithm is compared with some of the best performing graph matching methods on four datasets: simulated graphs, QAPLib, retina vessel images and handwritten chinese characters. In all cases, the results are competitive with the state-of-the-art [15].

6.4.4. A Tensor-Based framework for High-Order Graph Matching (O. Duchenne, F. Bach, J. Ponce, in collaboration with I. Kweon, KAIST university, South Corea)

We address the problem of establishing correspondences between two sets of visual features using higher-order constraints instead of the unary or pairwise ones used in classical methods. Concretely, the corresponding hypergraph matching problem is formulated as the maximization of a multilinear objective function over all permutations of the features. This function is defined by a tensor representing the affinity between feature tuples. It is maximized using a generalization of spectral techniques where a relaxed problem is first solved by a multi-dimensional power method, and the solution is then projected onto the closest assignment matrix [22]. In Figure 5, examples of matching two different objects from the same class of object.

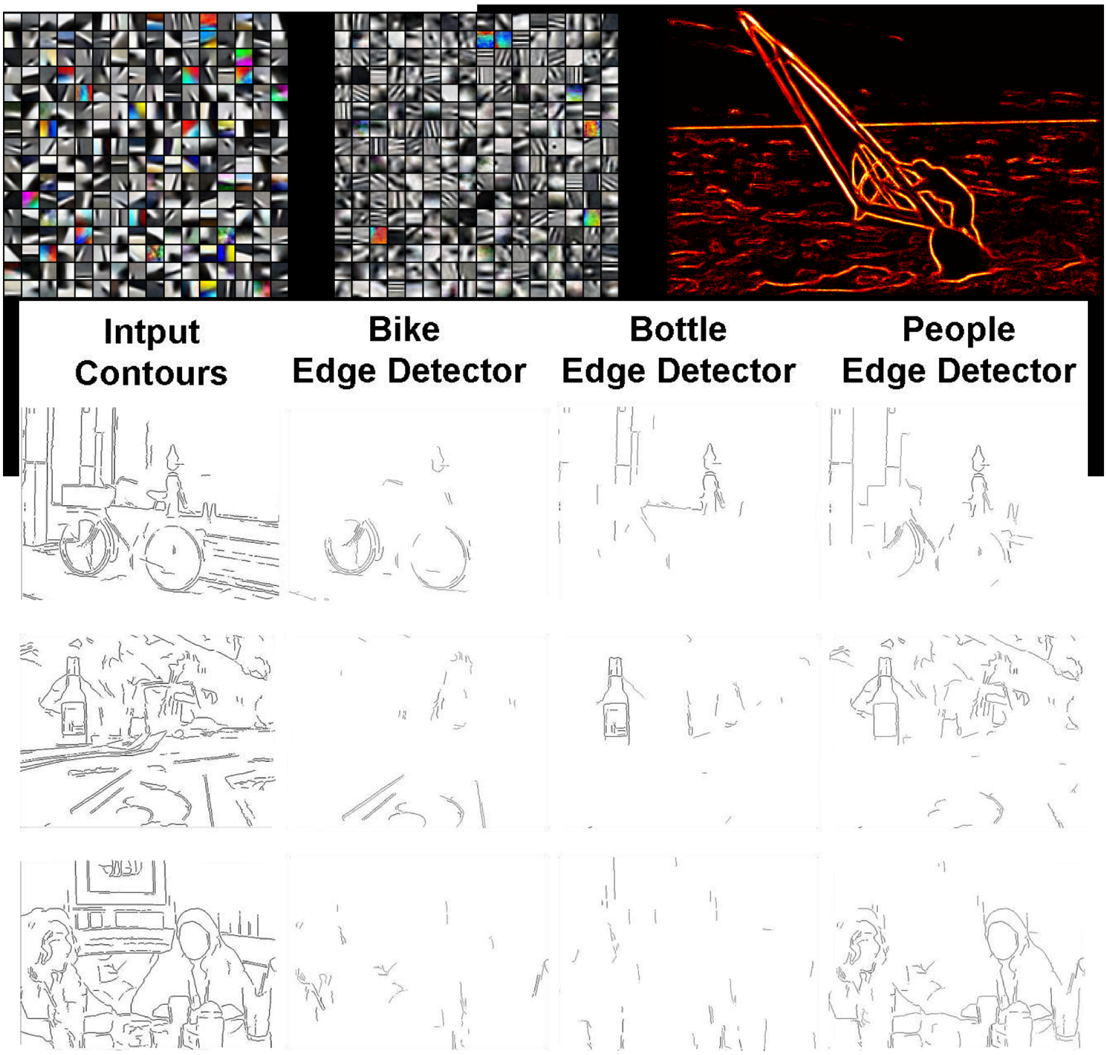

Figure 3. Top left: Two discriminative dictionaries, one for edges, one for background. Top right: An edge map computed by our algorithm. Bottom: Examples of class-specific edge detection for three classes, bike, bottle and people.

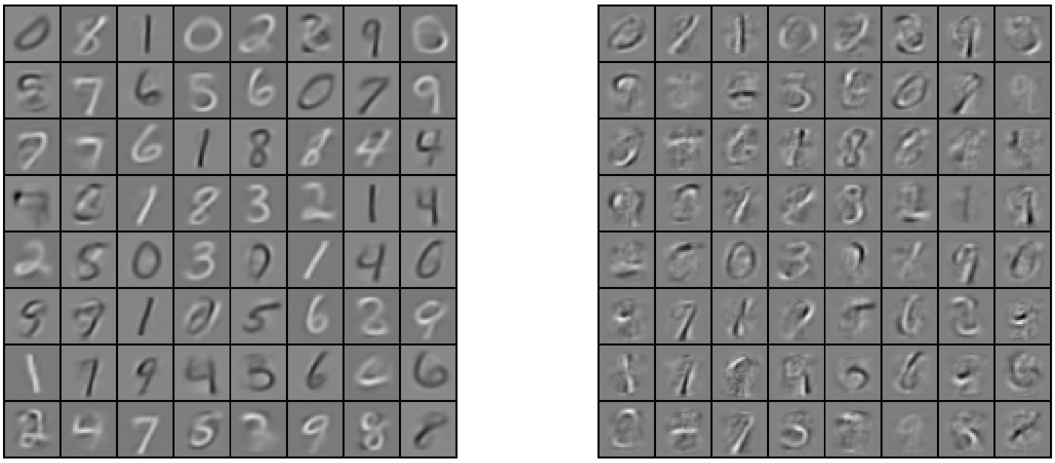

Figure 4. Left: Dictionary learned using the generative approach. Right: Dictionary learned using the discriminative approach

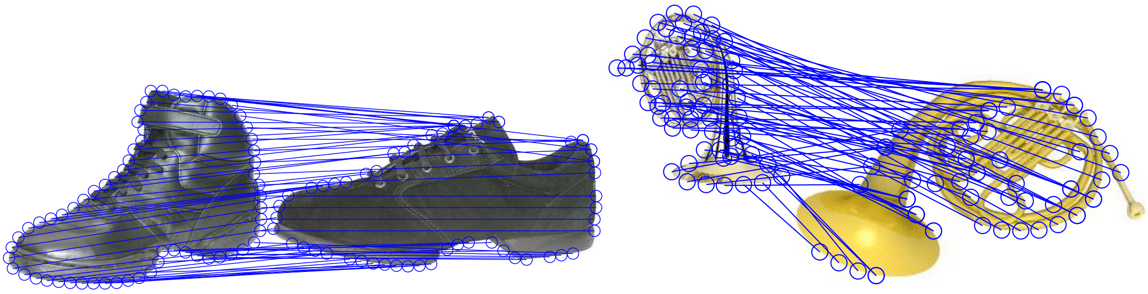

Figure 5. Matching silhouettes from the Caltech-256 database.

6.4.5. SIFT Flow: Dense Correspondence across Different Scenes (J. Sivic, joint work with C. Liu (MIT), J. Yuen (MIT), A. Torralba (MIT), and W.T. Freeman (MIT))

While image registration has been studied in different areas of computer vision, aligning images depicting different scenes remains a challenging problem, closer to recognition than to image matching. Analogous to optical flow, where an image is aligned to its temporally adjacent frame, we propose SIFT flow [28], a method to align an image to its neighbors in a large image collection consisting of a variety of scenes. For a query image, histogram intersection on a bag-of-visual-words representation is used to find the set of nearest neighbors in the database. The SIFT flow algorithm then consists of matching densely sampled SIFT features between the two images, while preserving spatial discontinuities. The use of SIFT features allows robust matching across different scene/object appearances and the discontinuity-preserving spatial model allows matching of objects located at different parts of the scene. Experiments show that the proposed approach is able to robustly align complicated scenes with large spatial distortions. We collect a large database of videos and apply the SIFT flow algorithm to two applications: (i) motion field prediction from a single static image and (ii) motion synthesis via transfer of moving objects. Figure 6 shows examples of synthesized motions for three different scenes.

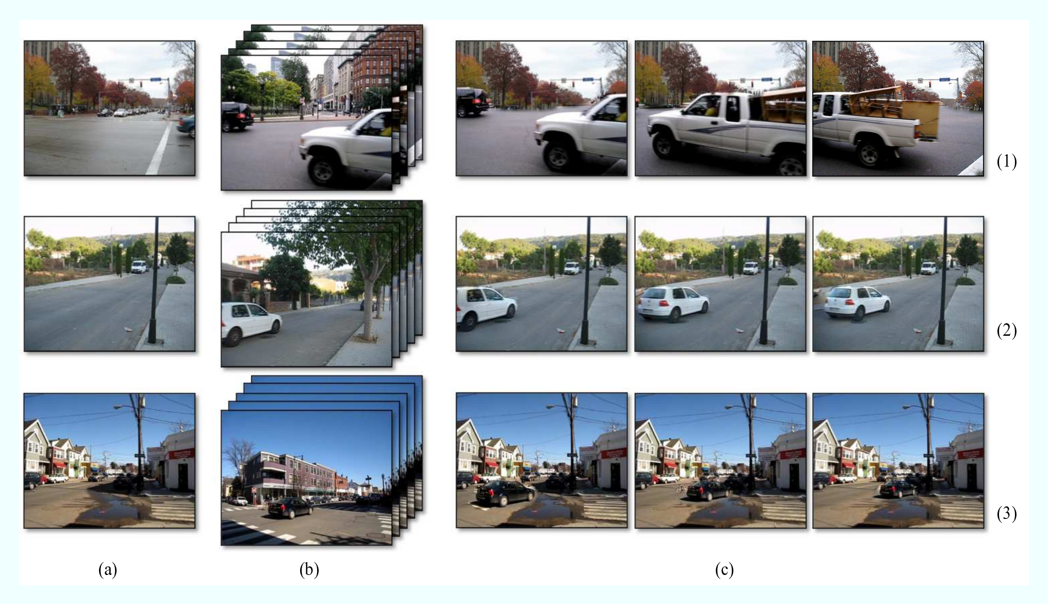

Figure 6. ***Category-level scene alignment applied to motion synthesis via object transfer.*** *Query still image (a), the top video match of a scene containing moving objects (b), and representative frames from the synthesized sequence (c) obtained by transferring moving objects by computing dense correspondence from the video to the still query image.*

6.4.6. A discriminative part model for object detection (J. Ponce, joint work with A. Kushal (UIUC) and C. Schmid (LEAR))

We have proposed in [45] a novel discriminative probabilistic framework for visual object category recognition where object classes are represented by assemblies of discriminative part models (or DPMs) obeying loose local geometric constraints. The parts are detected as parallelogram-shaped image regions and their appearance is modeled using a combination of a dense spatial pyramid histogram of gradient and edge intensities and a sparse spatial pyramid histogram of visual words. The parts are linked to enforce both geometric and cooccurrence constraints in a probabilistic graphical model. Geometric consistency is enforced locally among nearby object parts, which provides both an effective method for pruning mis-detections, and robustness to viewpoint changes and within-class shape variations. The appearance and geometric parameters are learned simultaneously and efficiently in a single convex optimization process. Our implementation yields state-of-the-art results on the PASCAL VOC Challenge 2007 car, cow and motorbike datasets (Figure 7).

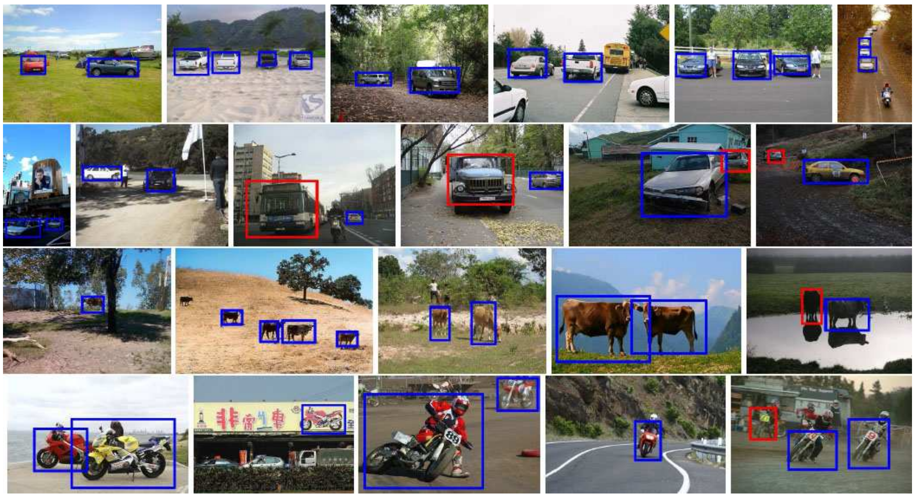

Figure 7. Object detection results on PASCAL VOC'07 cars, cows and motorbikes. The blue bounding boxes correspond to correct detections. The red boxes correspond to misdetections—some of these correspond to unlabeled objects (or those labeled as difficult) in the test set.

6.4.7. Unsupervised Discovery of Visual Object Class Hierarchies (J. Sivic, B. Russell and A. Zisserman, joint work with A. Efros (CMU) and W.T. Freeman (MIT))

Objects in the world can be arranged into a hierarchy based on their semantic meaning (e.g. organism - animal - feline - cat). What about defining a hierarchy based on the visual appearance of objects? This work [39] 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 non-overlapping 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 modeled using quantized local image regions as analogues to words in text. Employing the multiple segmentation framework of Russell et al., CVPR'06, we show that meaningful object hierarchies, together with object segmentations, can be automatically learned from unlabeled and unsegmented image collections without supervision. We demonstrate improved object classification and localization performance using hLDA over the previous non-hierarchical method on the MSRC dataset of Winn et al., ICCV'05.

6.5. Human activity capture and classification

6.5.1. Learning person specific classifiers from video (J. Sivic and A. Zisserman, joint work with M. Everingham (University of Leeds, UK))

We investigate the problem of automatically labelling appearances of characters in TV or movie material with their names, using only weak supervision in the form of automatically-aligned subtitle and script text. Previous work by Everingham *et al.*, BMVC 2006 demonstrated promising results on the task, but the coverage of the method (proportion of video labelled) and generalization was limited by a restriction to frontal faces and nearest neighbour classification.

We build in [37] on that method, extending the coverage greatly by the detection and recognition of characters in profile views. In addition, we make the following contributions: (i) seamless tracking and integration of profile and frontal detections, and (ii) a character specific multiple kernel classifier which is able to learn the features best able to discriminate between the characters.

We report results on two episodes of the TV series “Buffy the Vampire Slayer”, demonstrating significantly increased coverage and performance, with respect to previous methods on this material. Examples of correctly detected and named characters are shown in figure 8.

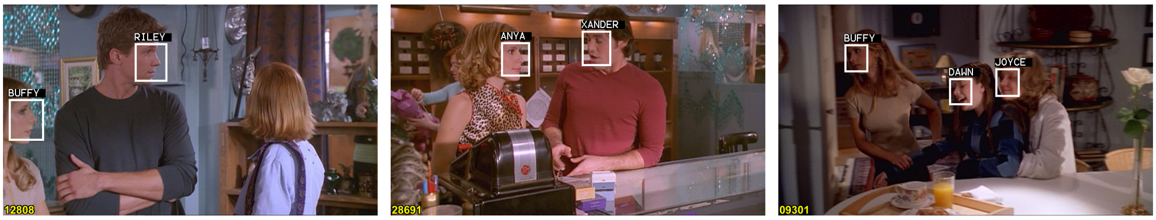

Figure 8. Examples of correct detection and naming of characters in non-frontal views from an episode of the TV series “Buffy the Vampire Slayer”.

6.6. Machine learning for computer vision

6.6.1. Robust image matching and recognition using context-dependent kernels (J.-Y. Audibert, joint work with R. Keriven, J. Rabarisoa and H. Sahbi)

The success of kernel methods including support vector machines (SVMs) strongly depends on the design of appropriate kernels. While initially kernels were designed in order to handle fixed-length data, their extension to unordered, variable-length data became more than necessary for real pattern recognition problems such as object recognition and bioinformatics. We focus in this paper on object recognition using a new type of kernel referred to as “context- dependent”, which allows to take into account geometric constraints when matching images (Figure 9). Objects, seen as constellations of local features (interest points, regions, etc.), are matched by minimizing an energy function mixing (1) a fidelity term which measures the quality of feature matching, (2) a neighborhood criteria which captures the object geometry and (3) a regularization term. We show that the minimizer of this energy is a “context-dependent” Mercer kernel. Experiments conducted on object recognition show that when plugging our kernel in SVMs, we clearly outperform SVMs with “context-free” kernels [36].

6.6.2. Transductive segmentation (O. Duchenne, J.-Y. Audibert and J. Ponce, joint work with R. Keriven and F. Ségonne)

In [21], we extend and publish the work done last year on this topic. We consider a multi-zone segmentation of a single image when user-supplied seeds are provided in each region. 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 that is tightly related to diffusion processes. Our segmentation is modeled as the task of finding matting coefficients for unclassified pixels given known matting coefficients of seed pixels. The resulting segmentation procedure is simple, fast, and accurate (Figure 10). Comparison with other methods on natural images databases are given.

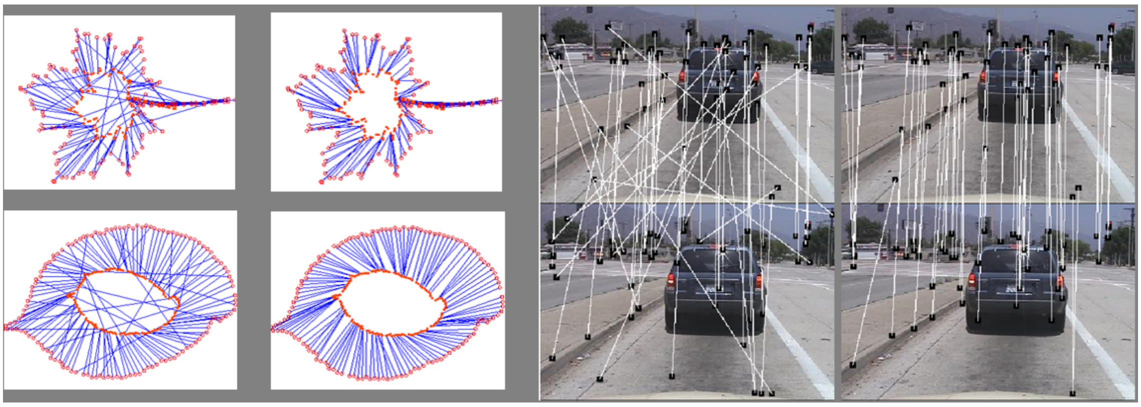

Figure 9. First and third columns: examples of matching results when using a naive matching strategy without geometry (two on leaf shape and one on a real-life image). Second and fourth columns: corresponding results with our "context-dependent" kernel.

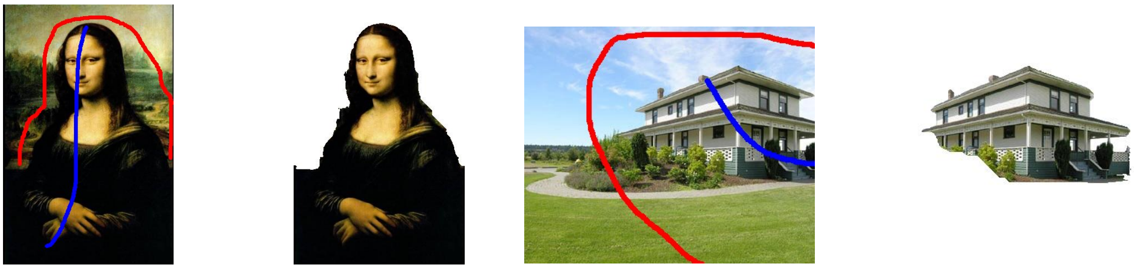

Figure 10. Results for our transductive segmentation method.

6.6.3. Semantic Lattices for Multiple Annotation of Images (J.-Y. Audibert, joint work with S. Herbin and A.-M. Tousch)

In [40], we address the problem of describing precisely an object present in an image. The starting point is a semantic lattice defining all possible coherent object descriptions through inheritance and exclusion relations. This domain knowledge is used in a learning process which outputs a set of coherent explanations of the image valued by their confidence level. Our first contribution is to design this method for multiple complexity level image description. Our secondary focus is to develop rigorous evaluation standards for this computer vision task which, to our knowledge, has not been addressed in the literature despite its possible use in symbolic annotation of multimedia database. A critical evaluation of our approach under the proposed standards is presented on a new appropriate car database that we have collected.

6.7. Effective learning algorithms and architectures

6.7.1. Scaling up machine learning algorithms by adaptive sampling (J.-Y. Audibert, joint work with V. Mnih and C. Szepesvári)

Sampling is a popular way of scaling up machine learning algorithms to large datasets and streaming data. The question often is how many samples are needed. Adaptive stopping algorithms monitor the performance in an online fashion and they can stop early, saving valuable resources. We consider problems where probabilistic guarantees are desired and demonstrate how recently-introduced empirical Bernstein bounds can be used to design stopping rules that are efficient. We provide upper bounds on the sample complexity of the new rules, as well as empirical results on model selection and boosting in the filtering setting [32].

6.7.2. Exploration-exploitation trade-off using variance estimates in multi-armed bandits (J.-Y. Audibert, joint work with R. Munos and C. Szepesvári)

Algorithms based on upper confidence bounds for balancing exploration and exploitation are gaining popularity since they are easy to implement, efficient and effective. In [5], we extend the work done last year on this topic. 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. We provide the first analysis of the expected regret for such algorithms. As expected, our results show that the algorithm that uses the variance estimates has a major advantage over its alternatives that do not use such estimates provided that the variances of the payoffs of the suboptimal arms are low. We also prove that the regret concentrates only at a polynomial rate. This holds for all the upper confidence bound based algorithms and for all bandit problems except those special ones where with probability one the payoff obtained by pulling the optimal arm is larger than the expected payoff for the second best arm. Hence, although upper confidence bound bandit algorithms achieve logarithmic expected regret rates, they might not be suitable for a risk-averse decision maker. We illustrate some of the results by computer simulations.

6.7.3. Algorithms for Infinitely Many-Armed Bandits (J.-Y. Audibert, joint work with R. Munos and Y. Wang)

Multi-armed bandit problems describe typical situations where learning and optimization should be balanced in order to achieve good cumulative performances. Usual multi-armed bandit problems consider a finite number of possible actions (or arms) from which the learner may choose at each iteration. The number of arms is typically much smaller than the number of experiments allowed, so exploration of all possible options is usually performed and combined with exploitation of the apparently best ones. In [41], we investigate the case when the number of arms is infinite (or larger than the available number of experiments), which makes the exploration of all the arms an impossible task to achieve: if no additional assumption is made, it may be arbitrarily hard to find a near-optimal arm. We make a stochastic assumption on the mean-reward of a new selected arm which characterizes its probability of being a near-optimal arm. Our assumption is weaker than in previous works. We describe algorithms based on upper-confidence-bounds applied to a restricted set of randomly selected arms and provide upper-bounds on the resulting expected regret. We also derive a lower-bound which matches (up to a logarithmic factor) the upper-bound in some cases.

6.8. Learning theory

6.8.1. Confidence regions and multiple testing by resampling (S. Arlot, joint work with G. Blanchard and É. Roquain)

Generalized bootstrapped confidence regions have been obtained in [2] for the mean of a random vector whose coordinates have an unknown dependence structure, with a non-asymptotic control of the confidence level. The random vector is supposed to be either Gaussian or to have a symmetric bounded distribution. We consider two approaches, the first one based on a concentration principle and the second one on a direct boostrapped quantile.

These results are applied in the one-sided and two-sided multiple testing problem [3], in which we derive several resampling-based step-down procedures providing a non-asymptotic FWER control. According to a simulation study, these procedures can outperform Bonferroni's or Holm's procedures as soon as the observed vector has sufficiently correlated coordinates.

6.8.2. Model selection by resampling (S. Arlot)

The classical V-fold cross-validation being biased, a penalization approach is proposed as an alternative, called V-fold penalization [42]. It can be used in a very general framework, and needs the same computation time as V-fold cross-validation. In the case example of regression on histograms, the V-fold penalties lead to a non-asymptotic oracle inequality, with constant almost one. This results holds with mild assumptions on the noise-level, showing that V-fold penalties are adaptive to heteroscedastic noises. Moreover, a simulation study shows that overpenalization may improve the quality of a model selection procedure, when the sample size is small, as compared to the noise level. The V-fold penalties allowing to choose separately V and the overpenalization factor, they are more flexible than V-fold cross-validation, and outperform it.

V-fold penalties have been generalized to a wide class of resampling penalties [43]. In the histogram regression case, a non-asymptotic oracle inequality and adaptation to the smoothness of the regression function and the heteroscedastic noise are proven. A simulation study in regression shows that resampling penalties outperform classical procedures such as Mallows' C_p and V-fold cross-validation.

6.8.3. Margin adaptive model selection in statistical learning (S. Arlot, joint work with P. L. Bartlett)

A classical condition for fast learning rates is the margin condition, first introduced by Mammen and Tsybakov. We tackle in [44] the problem of adaptivity to this condition in the context of model selection, in a general learning framework. Actually, we consider a weaker version of this condition that allows us to take into account that learning within a small model can be much easier than in a large one. Requiring this “strong margin adaptivity” makes the model selection problem more challenging. We first prove, in a very general framework, that some penalization procedures (including local Rademacher complexities) exhibit this adaptivity when the models are nested. Contrary to previous results, this holds with penalties that only depend on the data. Our second main result is that strong margin adaptivity is not always possible when the models are not nested: for every model selection procedure (even a randomized one), there is a problem for which it does not demonstrate strong margin adaptivity.

6.8.4. Fast learning rates in statistical inference through aggregation (J.-Y. Audibert)

In [4], we develop minimax optimal risk bounds for the general learning task consisting in predicting as well as the best function in a reference set $\mathcal{G}$ up to the smallest possible additive term, called the convergence rate. When the reference set is finite and when n denotes the size of the training data, we provide minimax convergence rates of the form $C\left(\frac{\log|\mathcal{G}|}{n}\right)^v$ with tight evaluation of the positive constant C and with exact $0 < v \leq 1$, the latter value depending on the convexity of the loss function and on the level of noise in the output distribution.

The risk upper bounds are based on a sequential randomized algorithm, which at each step concentrates on functions having both low risk and low variance with respect to the previous step prediction function. Our analysis puts forward the links between the probabilistic and worst-case viewpoints, and allows to obtain risk bounds unachievable with the standard statistical learning approach. One of the key idea of this work is to use probabilistic inequalities with respect to appropriate (Gibbs) distributions on the prediction function space instead of using them with respect to the distribution generating the data.

The risk lower bounds are based on refinements of the Assouad lemma taking particularly into account the properties of the loss function. Our key example to illustrate the upper and lower bounds is to consider the L_q-regression setting for which an exhaustive analysis of the convergence rates is given while q ranges in $[1;+\infty[$.

6.8.5. Exploring Large Feature Spaces with Hierarchical Multiple Kernel Learning (F. Bach)

For supervised and unsupervised learning, positive definite kernels allow to use large and potentially infinite dimensional feature spaces with a computational cost that only depends on the number of observations. This is usually done through the penalization of predictor functions by Euclidean or Hilbertian norms. We explore penalizing by sparsity-inducing norms such as the l1-norm or the block l1-norm. We assume that the kernel decomposes into a large sum of individual basis kernels which can be embedded in a directed acyclic graph; we show that it is then possible to perform kernel selection through a hierarchical multiple kernel learning framework, in polynomial time in the number of selected kernels. This framework is naturally applied to non linear variable selection; our extensive simulations on synthetic datasets and datasets from the UCI repository show that efficiently exploring the large feature space through sparsity-inducing norms leads to state-of-the-art predictive performance [19].

6.8.6. Kernel dimension reduction in regression (F. Bach, in collaboration with K. Fukumizu, Institute of Statistical Mathematics, Tokyo, and M. I. Jordan, U.C. Berkeley)

We design a new methodology for sufficient dimension reduction (SDR). Our methodology derives directly from the formulation of SDR in terms of the conditional independence of the covariate X from the response Y , given the projection of X on the central subspace. We show that this conditional independence assertion can be characterized in terms of conditional covariance operators on reproducing kernel Hilbert spaces and we show how this characterization leads to an M-estimator for the central subspace. The resulting estimator is shown to be consistent under weak conditions; in particular, we do not have to impose linearity or ellipticity conditions of the kinds that are generally invoked for SDR methods. We also present empirical results showing that the new methodology is competitive in practice [9].

6.8.7. Model consistent Lasso estimation through the bootstrap (F. Bach)

We consider the least-square linear regression problem with regularization by the L1-norm, a problem usually referred to as the Lasso. We perform a detailed asymptotic analysis of model consistency of the Lasso. For various decays of the regularization parameter, we compute asymptotic equivalents of the probability of correct model selection (i.e., variable selection). For a specific rate decay, we show that the Lasso selects all the variables that should enter the model with probability tending to one exponentially fast, while it selects all other variables with strictly positive probability. We show that this property implies that if we run the Lasso for several bootstrapped replications of a given sample, then intersecting the supports of the Lasso bootstrap estimates leads to consistent model selection. This novel variable selection algorithm, referred to as the Bolasso, is compared favorably to other linear regression methods on synthetic data and datasets from the UCI machine learning repository [18].

6.8.8. Kernel change-point analysis (F. Bach, in collaboration with Z. Harchaoui and Eric Moulines, Telecom Paris)

We introduce a kernel-based method for change-point analysis within a sequence of temporal observations. Change-point analysis of an (unlabelled) sample of observations consists in, first, testing whether a change in the distribution occurs within the sample, and second, if a change occurs, estimating the change-point instant after which the distribution of the observations switches from one distribution to another different distribution. We propose a test statistics based upon the maximum kernel Fisher discriminant ratio as a measure of homogeneity between segments. We derive its limiting distribution under the null hypothesis (no change occurs), and establish the consistency under the alternative hypothesis (a change occurs). This allows to build a statistical hypothesis testing procedure for testing the presence of change-point, with a prescribed false-alarm probability and detection probability tending to one in the large-sample setting. If a change actually occurs, the test statistics also yields an estimator of the change-point location. Promising experimental results in temporal segmentation of mental tasks from BCI data and pop song indexation are obtained [26].

6.8.9. SimpleMKL: efficient algorithms for multiple kernel learning (F. Bach, in collaboration with A. Rakotomamonjy and S. Canu, INSA Rouen, and Y. Grandvalet, UTC Compiègne)

Multiple kernel learning (MKL) aims at simultaneously learning a kernel and the associated predictor in supervised learning settings. For the support vector machine, an efficient and general multiple kernel learning algorithm, based on semi-infinite linear programming, has been recently proposed. This approach has opened new perspectives since it makes MKL tractable for large-scale problems, by iteratively using existing support vector machine code. However, it turns out that this iterative algorithm needs numerous iterations for converging towards a reasonable solution. We address the MKL problem through a weighted 2-norm regularization formulation with an additional constraint on the weights that encourages sparse kernel combinations. Apart from learning the combination, we solve a standard SVM optimization problem, where the kernel is defined as a linear combination of multiple kernels. We propose an algorithm, named SimpleMKL, for solving this MKL problem and provide a new insight on MKL algorithms based on mixed-norm regularization by showing that the two approaches are equivalent. We show how SimpleMKL can be applied beyond binary classification, for problems like regression, clustering (one-class classification) or multiclass classification. Experimental results show that the proposed algorithm converges rapidly and that its efficiency compares favorably to other MKL algorithms. Finally, we illustrate the usefulness of MKL for some regressors based on wavelet kernels and on some model selection problems related to multiclass classification problems [13].

6.8.10. A New Approach to Collaborative Filtering: Operator Estimation with Spectral Regularization (F. Bach, in collaboration with J. Abernethy, U.C. Berkeley, T. Evgeniou, INSEAD, and J.-P. Vert, Ecole des Mines de Paris)

We introduce a general approach for collaborative filtering (CF) using spectral regularization to learn linear operators from ÒusersÓ to the ÒobjectsÓ they rate. Recent low-rank type matrix completion approaches to CF are shown to be special cases. However, unlike existing regularization based CF methods, our approach can be used to also incorporate information such as attributes of the users or the objects—a limitation of existing regularization based CF methods. We then provide novel representer theorems that we use to develop new estimation methods. We provide learning algorithms based on low-rank decompositions, and test them on a standard CF dataset. The experiments indicate the advantages of generalizing the existing regularization based CF methods to incorporate related information about users and objects. Finally, we show that certain multi-task learning methods can be also seen as special cases of our proposed approach [1].

6.8.11. Clustered Multi-Task Learning (F. Bach, in collaboration with L. Jacob and J.-P. Vert, Ecole des Mines de Paris)

In multi-task learning several related tasks are considered simultaneously, with the hope that by an appropriate sharing of information across tasks, each task may benefit from the others. In the context of learning linear functions for supervised classification or regression, this can be achieved by including a priori information about the weight vectors associated with the tasks, and how they are expected to be related to each other. In this work, we assume that tasks are clustered into groups, which are unknown beforehand, and that tasks within a group have similar weight vectors. We design a new spectral norm that encodes this a priori assumption, without the prior knowledge of the partition of tasks into groups, resulting in a new convex optimization formulation for multi-task learning. We show in simulations on synthetic examples and on the IEDB MHC-I binding dataset, that our approach outperforms well-known convex methods for multi-task learning, as well as related non convex methods dedicated to the same problem [27].

6.8.12. Sparse probabilistic projections (F. Bach, in collaboration with C. Archambeau, University College London)

We consider a generative model for performing sparse probabilistic projections, which includes sparse principal component analysis and sparse canonical correlation analysis as special cases. Sparsity is enforced by means of automatic relevance determination or by imposing appropriate prior distributions, such as generalised hyperbolic distributions. We derive a variational Expectation-Maximisation algorithm for the estimation of the hyperparameters and show that our novel probabilistic approach compares favourably to existing techniques. The proposed method is applied in the context of cryptoanalysis as a preprocessing tool for the construction of template attacks [17].

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)

Participants: Jean Ponce, Jan van Gemert.

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). The effort in WILLOW is concerned with the detection of 3D targets and the estimation of their pose. Total WILLOW budget: 110 KEuros.

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

Participants: Jean-Yves Audibert, Jean Ponce, Hui Kong.

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, Josef Sivic, Andrew Zisserman.

A. Zisserman's participation in WILLOW has been partially funded by EADS. This has resulted in initial collaboration efforts via discussions and tutorial presentations by A. Zisserman, J. Ponce and J. Sivic at EADS. The tutorial was delivered at EADS Suresnes lab in November and December 2008 and covered efficient visual search of image and videos. In addition, Marc Sturzel (EADS) has started a PhD at ENS with Jean Ponce.

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: Jean Ponce, Josef Sivic, Oliwer Whyte, 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: Jean-Yves Audibert, Francis Bach, Olivier Duchenne, Julien Mairal, 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).

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. 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.

8.5. European Research Council (ERC) Senior Researcher grant

Participant: Andrew Zisserman.

Andrew Zisserman was awarded the European Research Council (ERC) Senior Researcher grant.

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). Organized paper submission, reviewing, Area Chair meeting in June and Conference Programme (http://eccv2008.inrialpes.fr/).
  - Organizer, joint Willow-Oxford workshop, April 2008 (A. Zisserman).
  - Co-organizer, Pascal VOC 2008 Workshop at the European Conference on Computer Vision, Marseille, 2008, http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/ (A. Zisserman).
- Editorial boards:
  - Journal of Machine Learning Research, Action Editor (F. Bach).
  - International Journal of Computer Vision (J. Ponce, A. Zisserman).
  - Foundations and Trends in Computer Graphics and Vision (J. Ponce).
- Area chairs:
  - Neural Information and Processing Systems (NIPS) Conference, 2008 (J.-Y. Audibert and F. Bach)
  - International Conference on Computer Vision, 2009 (J. Ponce)
- Program committees:
  - Conference on Learning Theory (COLT), 2008 (J.-Y. Audibert)
  - Conférence francophone sur l'apprentissage automatique (CAP), 2008 (J.-Y. Audibert)
  - IEEE Conference on Computer Vision and Pattern Recognition, 2007 (J. Ponce).
  - European Conference on Computer Vision, 2008 (J. Sivic).
  - British Machine Vision Conference, 2008 (J. Sivic).
  - Reviewer for Neural Information and Processing Systems (NIPS) Conference, 2008 (J. Sivic).
  - IEEE Conference on Computer Vision and Pattern Recognition, 2009 (J. Sivic).
- Prizes:
  - The paper *Geometric LDA: A Generative Model for Particular Object Discovery* by J. Philbin, J. Sivic and A. Zisserman was awarded the best poster prize at the British Machine Vision Conference (BMVC), 2008
- Other:

- J.-Y. Audibert is a member of the PASCAL2 European Network of Excellence (http://www.pascal-network.org).
- F. Bach is a member of the PASCAL2 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. 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/seminaires.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 PASCAL2 European Network of Excellence and co-organizes the Pascal VOC challenge (http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/).

9.2. Teaching

- J.-Y. Audibert, “Machine Learning and applications”, Ecole Nationale des Ponts et Chaussées, 2nd year, 21h.
- 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.
- Sparse methods for Machine learning, International Machine Learning Summer School, Ile de Re (F. Bach)
- J. Ponce, “Introduction to scientific computing”, Ecole normale supérieure, M1, 36h.
- J. Ponce, “Geometry and computer vision”, Ecole normale supérieure and MVA, Ecole normale supérieure de Cachan, 36h.
- J. Ponce and J. Sivic (together with C. Schmid (INRIA Grenoble)), “Object recognition and computer vision”, Ecole normale supérieure, and MVA, Ecole normale supérieure de Cachan, 36h.
- 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

- S. Arlot, *V-fold penalization: an alternative to V-fold cross-validation*, Cherry Bud Workshop, Keio University, Yokohama, Japan, March 2008.
- S. Arlot, *V-fold penalization: an alternative to V-fold cross-validation*, Second Canada-France Congress, UQAM, Montreal, Canada, June 2008.
- S. Arlot, *V-fold cross-validation improved: V-fold penalization*, Journées Statistiques du Sud, INSA Toulouse, June 2008.

- J.-Y. Audibert, *Transductive Learning and Computer Vision*, NIPS Workshop, New challenges in theoretical machine learning: learning with data-dependent concept spaces, Whistler, Canada, Dec. 2008
- J.-Y. Audibert, *Aggregation to compete with the best prediction function in a fixed set*, Swiss Probability Seminar, Bern, Switzerland, Nov. 2008
- J.-Y. Audibert, *Aggregation to compete with the best prediction function in a fixed set*, INRIA Sequel seminar, Lille, Nov. 2008
- J.-Y. Audibert and F. Bach, *Supervised Machine Learning*, Tutorial, ECCV, Marseille, Oct. 2008
- J.-Y. Audibert, *Graph Laplacian for transductive learning: application to image segmentation and interactive image search*, Rencontre “Modélisation Statistique des Images”, Luminy, May 2008
- J. Ponce, Ecole Polytechnique, Paris.
- J. Ponce, Microsoft Tech Days, Paris.
- J. Ponce, Télécom Paris.
- J. Ponce, European Workshop on Computational Geometry, Nancy.
- J. Ponce, International Workshop on Computer Vision, Venice.
- J. Ponce, University of Illinois at Urbana-Champaign.
- J. Ponce, University of California at Los Angeles.
- J. Ponce, University of Southern California.
- J. Ponce, International Workshop on Shape Perception in Human and Computer Vision, ECCV’08.
- J. Sivic, Harvard University, Cambridge, USA
- J. Sivic, Massachusetts Institute of Technology, Cambridge, USA
- J. Sivic, University of Washington, Seattle, USA
- J. Sivic, Microsoft Research, Redmond, USA
- J. Sivic, University of California, Berkeley, USA
- J. Sivic, École nationale supérieure des telecommunications, Paris, France
- J. Sivic, Scene Understanding Symposium, Massachusetts Institute of Technology, Cambridge, USA
- J. Sivic, Pattern Recognition and Computer Vision Colloquium, Czech Technical University in Prague
- J. Sivic, Joint Willow-Oxford workshop, University of Oxford, UK
- A. Zisserman, Key-note speaker at EU Network of Excellence (NOE) MUSCLE conference, February 2008, Cannes. http://www.muscle-noe.org/content/view/164/39/
- A. Zisserman, Invited talk at International Workshop on Computer Vision, Venice, May, 2008
- A. Zisserman, Speaker at the International Computer Vision Summer School 2008 (ICVSS 2008) http://svg.dmi.unict.it/icvss2008/

10. Bibliography

Year Publications

Articles in International Peer-Reviewed Journal

[1] J. Abernethy, F. Bach, T. Evgeniou, J.-P. Vert. *A New Approach to Collaborative Filtering: Operator Estimation with Spectral Regularization*, in "Journal of Machine Learning Research", to appear, 2008.

[2] S. Arlot, G. Blanchard, É. Roquain. *Some non-asymptotic results on resampling in high dimension, I: confidence regions*, in "Ann. Statist.", To appear, 2008.

[3] S. Arlot, G. Blanchard, É. Roquain. *Some non-asymptotic results on resampling in high dimension, II: multiple tests*, in "Ann. Statist.", To appear, 2008.

[4] J. Audibert. *Fast learning rates in statistical inference through aggregation*, in "Annals of Statistics", Jun 2008.

[5] J. Audibert, R. Munos, C. Szepesvari. *Exploration-exploitation trade-off using variance estimates in multi-armed bandits*, in "Theoretical Computer Science", 2008.

[6] F. Bach. *Consistency of the group Lasso and multiple kernel learning*, in "Journal of Machine Learning Research", vol. 9, 2008, p. 1179-1225.

[7] F. Bach. *Consistency of trace norm minimization*, in "Journal of Machine Learning Research", vol. 9, 2008, p. 1019-1048.

[8] M. Everingham, J. Sivic, A. Zisserman. *Taking the bite out of automatic naming of characters in TV video*, in "Image and Vision Computing", To appear, 2008.

[9] K. Fukumizu, F. Bach, M. I. Jordan. *Kernel dimension reduction in regression*, in "Annals of Statistics", to appear, 2008.

[10] Y. Furukawa, J. Ponce. *Accurate Camera Calibration from Multi-View Stereo and Bundle Adjustment*, in "Int. J. of Comp. Vision", Accepted with minor revision., 2008.

[11] Y. Furukawa, J. Ponce. *Carved Visual Hulls for Image-Based Modeling*, in "Int. J. of Comp. Vision", Published on-line, doi 10.1007/s11263-008-0134-8., 2008.

[12] Y. Furukawa, J. Ponce. *Accurate, Dense, and Robust Multi-View Stereopsis*, in "IEEE Trans. Patt. Anal. Mach. Intell.", Submitted., 2009.

[13] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet. *SimpleMKL*, in "Journal of Machine Learning Research", vol. 9, 2008, p. 2491–2521.

[14] J. Sivic, A. Zisserman. *Efficient Visual Search for Objects in Videos*, in "Proceedings of the IEEE", vol. 96, n^{o} 4, 2008, p. 548–566.

[15] M. Zaslavskiy, F. Bach, J.-P. Vert. *A path following algorithm for the graph matching problem*, in "IEEE Trans. Patt. Anal. Mach. Intell.", to appear, 2008.

[16] A. d'Aspremont, F. Bach, L. E. Ghaoui. *Optimal solutions for sparse principal component analysis*, in "Journal of Machine Learning Research", vol. 9, 2008, p. 1269-1294.

International Peer-Reviewed Conference/Proceedings

[17] C. Archambeau, F. Bach. *Sparse probabilistic projections*, in "Proc. Neural Info. Proc. Systems", 2008.

[18] F. BACH. *Bolasso: Model consistent Lasso estimation through the bootstrap*, in "Proc. Int. Conf. on Machine Learning", 2008.

[19] F. BACH. *Exploring Large Feature Spaces with Hierarchical Multiple Kernel Learning*, in "Proc. Neural Info. Proc. Systems", 2008.

[20] F. BACH. *Graph kernels between point clouds*, in "Proc. Int. Conf. on Machine Learning", 2008.

[21] O. DUCHENNE, J. AUDIBERT, R. KERIVEN, J. PONCE, F. SÉGONNE. *Segmentation by transduction*, in "Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, Alaska", Jun 2008.

[22] O. DUCHENNE, F. BACH, J. PONCE, I. KWEON. *A Tensor-Based algorithm for High-Order Graph Matching*, in "Proc. IEEE Conf. Comp. Vision Patt. Recog.", submitted, 2009.

[23] Y. FURUKAWA, J. PONCE. *Accurate Camera Calibration from Multi-View Stereo and Bundle Adjustment*, in "Proc. IEEE Conf. Comp. Vision Patt. Recog.", 2008.

[24] Y. FURUKAWA, J. PONCE. *Dense 3D Motion Capture from Synchronized Video Streams*, in "Proc. IEEE Conf. Comp. Vision Patt. Recog.", 2008.

[25] Y. FURUKAWA, J. PONCE. *Dense 3D Motion Capture for Faces*, in "Proc. IEEE Conf. Comp. Vision Patt. Recog.", Submitted., 2009.

[26] Z. HARCHAOUI, F. BACH, E. MOULINES. *Kernel change-point analysis*, in "Proc. Neural Info. Proc. Systems", 2008.

[27] L. JACOB, F. BACH, J.-P. VERT. *Clustered Multi-Task Learning: A Convex Formulation*, in "Proc. Neural Info. Proc. Systems", 2008.

[28] C. LIU, J. YUEN, A. TORRALBA, J. SIVIC, W. T. FREEMAN. *SIFT Flow: Dense Correspondence across Different Scenes*, in "Proceedings of the 10th European Conference on Computer Vision, Marseille, France", October 2008.

[29] J. MAIRAL, F. BACH, J. PONCE, G. SAPIRO, A. ZISSERMAN. *Discriminative Learned Dictionaries for Local Image Analysis*, in "Proc. IEEE Conf. Comp. Vision Patt. Recog.", 2008.

[30] J. MAIRAL, F. BACH, J. PONCE, G. SAPIRO, A. ZISSERMAN. *Supervised Dictionary Learning*, in "Proc. Neural Info. Proc. Systems", 2008.

[31] J. MAIRAL, M. LEORDEANU, F. BACH, M. HEBERT, J. PONCE. *Discriminative Sparse Image Models for Class-Specific Edge Detection and Image Interpretation*, in "Proc. European Conf. Comp. Vision", 2008.

[32] V. MNIH, C. SZEPESVARI, J. AUDIBERT. *Empirical Bernstein stopping*, in "International Conference on Machine Learning (ICML), Helsinki, Finlande", Jul 2008.

[33] J. PHILBIN, O. CHUM, M. ISARD, J. SIVIC, A. ZISSERMAN. *Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", 2008.

[34] J. PHILBIN, J. SIVIC, A. ZISSERMAN. *Geometric LDA: A Generative Model for Particular Object Discovery*, in "Proceedings of the British Machine Vision Conference", 2008.

[35] J. PONCE. *What is a Camera?*, in "Proc. IEEE Conf. Comp. Vision Patt. Recog.", Submitted., 2009.

[36] H. SAHBI, J. AUDIBERT, J. RABARISOA, R. KERIVEN. *Robust Matching and Recognition using Context-Dependent Kernels*, in "25th International Conference on Machine Learning (ICML), Helsinki, Finlande", Jul 2008.

[37] J. SIVIC, M. EVERINGHAM, A. ZISSERMAN. *"Who are you?": Learning person specific classifiers from video*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", submitted, 2009.

[38] J. SIVIC, B. KANEVA, A. TORRALBA, S. AVIDAN, W. T. FREEMAN. *Creating and Exploring a Large Photorealistic Virtual Space*, in "Proceedings of the First IEEE Workshop on Internet Vision, Ancorage, Alaska, USA", June 2008.

[39] J. SIVIC, B. C. RUSSELL, A. ZISSERMAN, W. T. FREEMAN, A. A. EFROS. *Unsupervised Discovery of Visual Object Class Hierarchies*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", 2008.

[40] A. TOUSCH, S. HERBIN, J. AUDIBERT. *Semantic Lattices for Multiple Annotation of Images*, in "ACM International Conference on Multimedia Information Retrieval (MIR), Vancouver, Canada", Oct 2008.

[41] Y. WANG, J. AUDIBERT, R. MUNOS. *Algorithms for Infinitely Many-Armed Bandits*, in "Advances in Neural Information Processing Systems, Vancouver, Canada", Dec 2008.

Research Reports

[42] S. ARLOT. *V-fold cross-validation improved: V-fold penalization*, Technical report, arXiv:0802.0566, 2008, http://fr.arxiv.org/abs/0802.0566.

[43] S. ARLOT. *Model selection by resampling penalization*, Technical report, hal-00262478, 2008, http://hal.archives-ouvertes.fr/hal-00262478/en/.

[44] S. ARLOT, P. L. BARTLETT. *Margin adaptive model selection in statistical learning*, Technical report, arXiv:0804.2937, 2008, http://fr.arxiv.org/abs/0804.2937.

[45] A. KUSHAL, C. SCHMID, J. PONCE. *A discriminative part model for object detection*, Technical report, 2008.