INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team WILLOW

Models of Visual Object Recognition and Scene Understanding

Paris - Rocquencourt

Theme : Vision, Perception and Multimedia Understanding

Activity Report

2009

Table of contents

1. **Team** 1
2. **Overall Objectives** 2
3. **Scientific Foundations** 2
   3.1. 3D object and scene modeling, analysis, and retrieval 2
      3.1.1. High-fidelity image-based object and scene modeling. 3
      3.1.2. Video-based modeling of deformable surfaces. 3
      3.1.3. Retrieval and modeling of objects and scenes in large image collections. 3
   3.2. Category-level object and scene recognition 3
      3.2.1. Learning image and object models. 3
      3.2.2. Category-level object/scene recognition and segmentation 4
   3.3. Human activity capture and classification 4
      3.3.1. Naming and recognition of characters in TV video 4
      3.3.2. Weakly-supervised learning and annotation of human actions in video 4
      3.3.3. Descriptors for video representation 4
   3.4. Machine learning 4
      3.4.1. Machine learning for computer vision. 4
      3.4.2. Algorithms and Learning theory. 5
4. **Application Domains** 5
   4.1. Introduction 5
   4.2. Quantitative image analysis in science and humanities 5
   4.3. Film Post-Production and Special Effects 5
   4.4. Video Annotation, Interpretation, and Retrieval 5
5. **Software** 6
   5.1. SPAMS 6
   5.2. Hierarchical kernel learning 6
   5.3. Structured variable selection 6
   5.4. Confidence regions and multiple tests by resampling 6
6. **New Results** 6
   6.1. High-fidelity image- and video-based modeling 6
      6.1.1. What is a camera? (J.Ponce, joint work with Guillaume Batog and Xavier Goaoc, VEGAS project-team) 6
      6.1.2. Quantitative image analysis for archeology (B. Russell, J. Ponce, joint work with H. Dessales, ENS Archeology laboratory) 7
      6.1.3. Dense 3D motion capture for human faces (J. Ponce, joint work with Y. Furukawa, University of Washington) 8
      6.1.4. Webcam Clip Art (A. Efros, joint work with J.-F. Lalonde, and S. G. Narasimhan, CMU) 8
   6.2. Retrieval and modeling of objects and scenes in large image collections 9
      6.2.1. Geometric Latent Dirichlet Allocation on a Matching Graph for Large-Scale Image Datasets (J. Sivic and A. Zisserman, joint work with J. Philbin, Oxford University) 9
      6.2.2. Infinite Images: 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) 9
      6.2.3. Get out of my picture! Internet based inpainting (O. Whyte, J. Sivic and A. Zisserman) 10
      6.2.4. Avoiding confusing features in place recognition (J. Sivic, in collaboration with J. Knopp, CTU Prague / KU Lueven, and T. Pajdla, CTU Prague) 10
      6.2.5. Looking beyond image boundaries (J. Sivic, in collaboration with B. Kaneva, MIT, S. Avidan, Adobe, W. T. Freeman, MIT, and A. Torralba, MIT) 11
   6.3. Learning image and object models 11

6.3.1. Non-local Sparse Models for Image Restoration (J. Mairal, F. Bach, J. Ponce, joint work with G. Sapiro, University of Minnesota) 11
6.3.2. Learning mid-level features for recognition (Y.-L. Boureau and J. Ponce, joint work with Y. LeCun, New York University) 11
6.3.3. Reasoning About Object Relationships (A. Efros, with T. Malisiewicz, CMU) 13
6.3.4. Non-uniform Deblurring for Shaken Images (O. Whyte, J. Sivic, A. Zisserman and J. Ponce) 13
6.3.5. Segmenting Scenes by Matching Image Composites (B. Russell, A. Efros, J. Sivic and A. Zisserman, joint work with W. T. Freeman, MIT) 13
6.3.6. Learning from Ambiguously Labeled Images (T. Cour, B. Sapp, C. Jordan and B. Taskar) 13
6.3.7. Learning discriminative part-based object models from weakly annotated images (T. Cour and F. Bach) 14
6.4. Human activity capture and classification 14
6.4.1. Automatic Annotation of Human Actions in Video (O. Duchenne, I. Laptev, J. Sivic, F. Bach and J. Ponce) 14
6.4.2. Actions in Context (I. Laptev, joint work with M. Marszałek and C. Schmid, INRIA-Grenoble) 14
6.4.3. Evaluation of local spatio-temporal features for action recognition (I. Laptev, joint work with M. M. Ullah at INRIA-Rennes and H. Wang, A. Kläser C. Schmid at INRIA-Grenoble) 15
6.4.4. Multi-view Synchronization of Human Actions and Dynamic Scenes (I. Laptev, joint work with E. Dexter and P. Pérez at INRIA-Rennes) 16
6.4.5. Quantitative analysis of videos for social sciences (N. Cherniavsky, I. Laptev, J. Ponce, J. Sivic, A. Zisserman) 16
6.4.6. Learning person specific classifiers from video (J. Sivic and A. Zisserman, joint work with M. Everingham, University of Leeds) 17
6.5. Machine learning for computer vision 17
6.5.1. Online Matrix Factorization for Sparse Coding (J. Mairal, F. Bach, J. Ponce, joint work with G. Sapiro, University of Minnesota) 17
6.5.2. Vanishing point detection for road detection (H. Kong, J.-Y. Audibert and J. Ponce) 17
6.5.3. Transductive segmentation of textured meshes (J.-Y. Audibert, joint work with A.-L. Jachiet, J.-P. Pons and R. Keriven) 17
6.6. Effective learning algorithms and architectures 19
6.6.1. Structured variable selection with sparsity-inducing norms (R. Jenatton, J-Y. Audibert and F. Bach) 19
6.6.2. Structured sparse principal component analysis (R. Jenatton, G. Obozinski and F. Bach) 20
6.6.3. Group Lasso with Overlap and Graph Lasso (G. Obozinksi, joint work with L. Jacob and J.-P. Vert, Ecole des Mines de Paris 20
6.6.4. Global alignment of protein interaction networks by graph matching methods (F. Bach, in collaboration with M. Zaslavskiy and J.-P. Vert, Ecole des Mines de Paris) 20
6.7. Learning theory 21
6.7.1. Data-driven calibration of linear estimators with minimal penalties (S. Arlot and F. Bach) 21
6.7.2. Resampling-based estimation of the accuracy of satellite ephemerides (S. Arlot, joint work with J. Desmars, J.-E. Arlot, V. Lainey and A. Vienne) 22
6.7.3. Asymptotically optimal regularization in smooth parametric models. (F. Bach, with P. Liang, G. Bouchard, M. I. Jordan) 22
6.7.4. Minimax policies for adversarial and stochastic bandits (J.-Y. Audibert) 22
6.7.5. Linear least squares regression (J.-Y. Audibert) 22
6.7.6. Change-point detection (S. Arlot, joint work with A. Celisse) 22

7. Contracts and Grants with Industry 23
7.1. Introduction 23
7.2. DGA/Bertin/EADS/SAGEM: 2ACI (ENS) 23
7.3. DGA/E-vitech: ITISECURE (ENS) 23
7.4. EADS (ENS) 23
7.5. MSR-INRIA joint lab: Image and video mining for science and humanities (INRIA) 23
8. Other Grants and Activities 24
8.1. Agence Nationale de la Recherche: HFIMBR (INRIA) 24
8.2. Agence Nationale de la Recherche: DETECT (ENS) 24
8.3. Agence Nationale de la Recherche: MGA (INRIA/ENPC) 24
8.4. Agence Nationale de la Recherche: Triangles (ENS) 24
8.5. France-UC Berkeley fund (Ecole des Mines de Paris) 25
8.6. European Research Council (ERC) Starting Investigator Researcher grant 25
9. Dissemination 25
9.1. Leadership within the scientific community 25
9.2. Teaching 26
9.3. Invited presentations 27
9.4. Popular science 28
10. Bibliography 28

Willow is a common project with the Ecole Normale Supérieure de Paris. The team has been created on January, 1^{st}, 2007 and became an INRIA project on June, 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, HdR]
Ivan Laptev [Chargé de Recherches INRIA]
Josef Sivic [Chargé de Recherches INRIA]

PhD Student

Louise Benoit
Y-Lan Boureau
Florent Couzinié-Devy
Olivier Duchenne
Loic Février
Toby Hocking
Rodolphe Jenatton
Armand Joulin
Augustin Lefèvre
Julien Mairal
Marc Sturzel
Oliver Whyte

Post-Doctoral Fellow

Neva Cherniavsky
Timothée Cour
Hui Kong
Guillaume Obozinski
Bryan Russell
Johannes van Gemert

Visiting Scientist

Léon Bottou [Research scientist with NEC Labs, Princeton, USA]
Frédo Durand [Associate professor in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology]
Alexei Efros [Assistant Professor in the Robotics Institute and Computer Science Department of the Carnegie Mellon University]
Ramakant Nevatia [Professor in Computer Science and Electrical Engineering at the University of Southern California]

Administrative Assistant

Cécile Espiègle

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 one new researcher: Ivan Laptev ("chargé de recherche", INRIA) has joined WILLOW in July 2009, coming from the VISTA project-team (INRIA Rennes). In addition, we have hired three post-docs: Guillaume Obozinski, Neva Cherniavsky, and Timothée Cour, and five new PhD students: Armand Joulin, Augustin Lefèvre, Louise Benoit, Toby Hocking, Florent Couzinié-Devy. Alexei Efros (Professor, Carnegie Mellon University, USA), Frédo Durand (MIT), Léon Bottou (NEC, USA), and Ram Nevatia (USC, USA) visited WILLOW in 2009.

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 multi-view stereopsis algorithms 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://grail.cs.washington.edu/software/pmvs/) is available for free for academics, and licensing negotiations with several companies are under way. Our current work extends this approach in two directions: the first one is theoretical, with a general formalism for modeling central and non-central cameras using the formalism and terminology of classical projective geometry (Section 6.1.1), while the second one is more applied, using our multi-view-stereo approach to model archaeological sites (Section 6.1.2).

3.1.2. Video-based modeling of deformable surfaces.

Another focus of our research is markerless motion capture from multiple video streamsn. Our previous work was targetted at the acquisition of accurate models of the shape of motion of deformable surfaces that bend but do not stretch (for example, many types of cloth). Our current work is aimed at extending this approach to surfaces that may stretch or shrink, such as human skin (Section 6.1.3). The targetted application is performance capture in the film industry.

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

The goal of this research is to develop techniques for visual search and recognition of objects and scenes in large image collections. In addition, the goal is to also investigate novel applications of large scale recognition in other domains, such as image processing (e.g. image enhancement and restoration), computer graphics (novel scene synthesis, visualization), 3D reconstruction, or visual localization.

We have introduced a geometric Latent Dirichlet Allocation (gLDA) model for unsupervised modeling of unstructured image collections and developed an approach for avoiding confusing features (such as trees or road markings) in the context of large scale place recognition in structured, geo-referenced image databases. In terms of applications, we have (i) developed a method for image inpainting using strong priors in the form of multiple other images of the same scene and (ii) investigated scene category recognition techniques for synthesizing novel scenes and navigating 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.

Learning sparse representations of images has been the topic of much recent research. It has been used for instance for image restoration (e.g., Mairal et al., 2007) and it has been generalized to discriminative image understanding tasks such as texture segmentation, category-level edge selection and image classification (Mairal et al., 2008). As discussed in Section 6.5.1, we have developed fast and scalable optimization methods for learning the sparse image representations [37], [15] and developed a software called SPAMS (SPArse Modelling Software) presented in Section 5.1. We have also unified this framework for image with the so-called non-local means approach, which exploits image self-similarities, leading to state-of-the-art results for image denoising and image demosaicking [38]. We present this work in Section 6.3.1.

3.2.2. Category-level object/scene recognition and segmentation

Another signiÞcant strand of our research has focused on the extremely challenging goals of category-level object/scene recognition and segmentation. Towards these goals, we have developed new models or algorithms for (i) reasoning about object relationships in an image, (ii) scene segmentation based on data-driven boundary detection, (iii) learning mid-level features for recognition and (iv) learning object/part models from weakly or ambiguously annotated images.

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, Marszałek, Schmid, Rozenfeld, 2008) 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.

3.3.2. Weakly-supervised learning and annotation of human actions in video

We aim to leverage the huge amount of video data using readily-available annotations in the form of video scripts. Scripts, however, often provide only imprecise and incomplete information about the video. We address this problem with weakly-supervised learning techniques both at the text and image levels. To this end we recently explored automatic mining of scene categories and action-scene correlations and demonstrated advantage thereof when recognizing human actions and scenes in video. We also developed a discriminative clustering approach for human actions addressing imprecision in the temporal script-based video annotation.

3.3.3. Descriptors for video representation

Video representation has a crucial role for recognizing human actions and other components of a visual scene. Our work in this domain aims to develop generic methods for representing video data that rely on realistic assumptions only. We are studying different ways for representing shape and motion information, we also investigate view-stable representations for human actions.

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 specialized unsupervised matrix factorization algorithms for image representation and image denoising, 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. Algorithms and 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. We are currently exploring structured sparse methods, where the idea is to introduce some prior knowledge into a sparse inference problem, for computational reasons or to improve interpretability and predictive performance. Moreover, all these methods lead to interesting practical hyperparameter selection, which can be tackled by theoretically grounded data-driven calibration procedures.

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. SPAMS

SPAMS (Sparse Modeling Software) is an optimization toolbox for Matlab. It is composed of a set of binaries implementing state-of-the-art algorithms to address various machine learning and signal processing problems involving a large number of small/medium size sparse decompositions.

The software was registered at the French Agency for Protecting Programs (Agence pour la Protection des Programmes) under number IDDN.FR.001.380004.000.S.P.2009.000.31235 on 15 September 2009, and is freely available at http://www.di.ens.fr/willow/SPAMS/.

5.2. Hierarchical kernel learning

The HKL sotware is a Matlab package allowing to perform hierarchical kernel learning, which is a recent technique designed to perform non linear variable selection in regression and classification problems. The package is provided free for non-commercial use under the terms of the GNU General Public License. It is publicly available at http://www.di.ens.fr/~fbach/hkl.

5.3. Structured variable selection

The work [50] comes with its Matlab package allowing to perform structured variable selection and to solve structured regression tasks. The package is provided free for non-commercial use under the terms of the GNU General Public License. It is publicly available at http://www.di.ens.fr/~jenatton/.

5.4. Confidence regions and multiple tests by resampling

A family of resampling-based confidence regions and multiple testing procedures was properly defined and theoretically studied in the papers [3], [4]. This software is a Matlab package allowing to compute these confidence regions and to perform these multiple testing procedures. This package is provided free for non-commercial use under the terms of the GNU General Public License. It is publicly available at url http://www.di.ens.fr/~arlot/code/CRMTR.htm.

6. New Results

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

6.1.1. What is a camera? (J.Ponce, joint work with Guillaume Batog and Xavier Goaoc, VEGAS project-team)

We address in [41] 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.

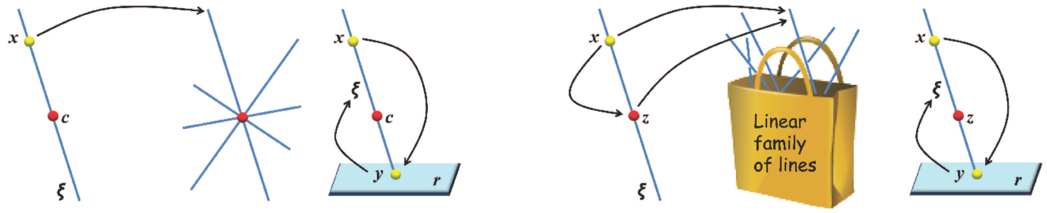


Figure 1. A pinhole camera can be thought of as a device that associates with any point $\mathbf{x}$ *the ray* ξ *that joins it to its image and passes through the pinhole* $\mathbf{c}$*. This ray is picked from the bundle of lines passing through* $\mathbf{c}$*. More generally, a (non-central) camera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

In [23], we extend this approach by presenting a complete analytical characterization of linear cameras: Pajdla has shown that a subset of these, the oblique cameras, can be modelled by a certain type of linear map. We have obtained a full tabulation of all admissible maps that induce cameras in the general sense of Grossberg and Nayar, and shown that these cameras are exactly the linear ones. Combining these two models with a new notion of intrinsic parameters and normalized coordinates for linear cameras allows us to give simplified analytical formulas for direct and inverse projections. We also show that the epipolar geometry of any two linear cameras can be characterized by a fundamental matrix whose size is at most 6×6 when the cameras are uncalibrated, or by an essential matrix of size at most 4×4 when their internal parameters are known. Similar results hold for trinocular constraints.

6.1.2. Quantitative image analysis for archeology (B. Russell, J. Ponce, joint work with H. Dessales, ENS Archeology laboratory)

Accurate indexing and alignment of images is an important problem in computer vision. A successful system would allow a user to retrieve images with similar content to a query image, along with any information associated with the image. Prior work has mostly focused on techniques to index and match photographs depicting particular instances of objects or scenes (e.g. famous landmarks, commercial product labels, etc.). This has allowed progress on tasks, such as the recovery of a 3D reconstruction of the depicted scene.

However, there are many types of images that cannot be accurately aligned. For instance, for many locations there are drawings and paintings made by artists that depict the scene. Matching and aligning photographs, paintings, and drawings is extremely difficult due to various distortions that can arise. Examples include perspective and caricature distortions, along with errors that arise due to the difficulty of drawing a scene by hand.

In this project, we seek to index and align a database of images, paintings, and drawings. The focus of our work is the Championnet house in the Roman ruins at Pompeii, Italy. Given an alignment of the images, paintings, and drawings, we wish to explore tasks that are of interest to archaeologists and curators who wish to study and preserve the site. Example applications include: (i) digitally restoring paintings on walls where the paintings have disappeared over time due to erosion, (ii) geometrically reasoning about the site over time through the drawings, (iii) indexing and searching patterns that exist throughout the site.

To date, we have visited the site in Pompeii and photographed the rooms of interest. An initial dense 3D reconstruction has been achieved from 585 photographs using existing photometric multi-view stereo methods. Figure 2 shows a snapshot of a 3D reconstruction of one of the rooms of interest. Notice that the 3D reconstruction captures much detail of the walls and structures.

We are currently exploring different techniques to align the photographs, paintings, and drawings. We hope to submit results from our research to a conference in Spring 2010.

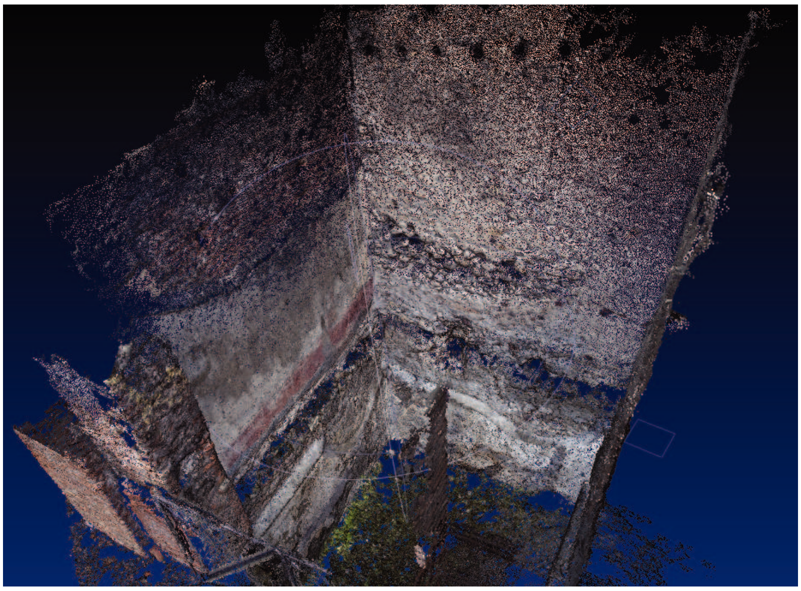


Figure 2. An initial dense 3D reconstruction of a room from the Championnet house in the Roman ruins at Pompeii, Italy. The reconstruction was computed from 585 photographs using existing photometric multi-view stereo methods. Notice that the reconstruction captures much detail of the walls and structures.

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

We have proposed in [30] 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.

6.1.4. Webcam Clip Art (A. Efros, joint work with J.-F. Lalonde, and S. G. Narasimhan, CMU)

Webcams placed all over the world observe and record the visual appearance of a variety of outdoor scenes over long periods of time. The recorded time-lapse image sequences cover a wide range of illumination and weather conditions – a vast untapped resource for creating visual realism. In this work, we propose to use a large repository of webcams as a "clip art" library from which users may transfer scene appearance (objects, scene backdrops, outdoor illumination) into their own time-lapse sequences or even single photographs. The goal is to combine the recent ideas from data-driven appearance transfer techniques with a general and theoretically-grounded physically-based illumination model. To accomplish this, the paper presents three main research contributions: 1) a new, high-quality outdoor webcam database that has been calibrated radiometrically and geometrically; 2) a novel approach for matching illuminations across different scenes based on the estimation of the properties of natural illuminants (sun, sky, weather and clouds), the camera geometry, and

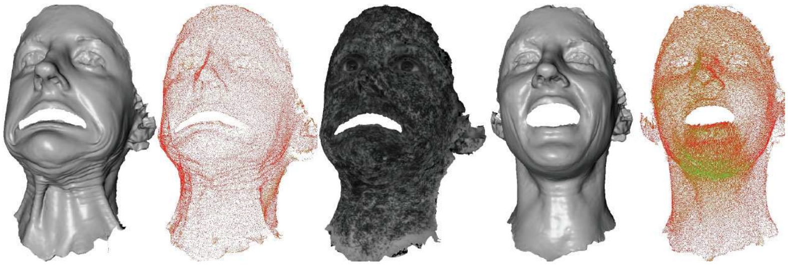


Figure 3. Facial motion capture [30], featuring shaded renderings of reconstructions obtained from two different frames, the corresponding dense motion fields, and one texture-mapped rendering (the actress's face was covered with make-up to provide additional texture). See http://www.cs.washington.edu/homes/furukawa/gallery/ for videos. Data courtesy of Image Movers Digital.

illumination-dependent scene features; 3) a new algorithm for generating physically plausible high dynamic range environment maps for each frame in a webcam sequence.

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

6.2.1. Geometric Latent Dirichlet Allocation on a Matching Graph for Large-Scale Image Datasets (J. Sivic and A. Zisserman, joint work with J. Philbin, Oxford University)

Given a large-scale collection of images we would like to be able to conceptually group together images taken of the same place, of the same thing, or of the same person.

To achieve this, we introduce the Geometric Latent Dirichlet Allocation (gLDA) model 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.

Additionally, to reduce the computational cost of applying our model to large datasets, we describe a scalable method that first computes a matching graph over all the images in a dataset. This matching graph connects images that contain the same object and rough image groups can be mined from this graph using standard clustering techniques. The gLDA model can then be applied to generate a more nuanced representation of the data. We also discuss how “hub images” (images representative of an object or landmark) can easily be extracted from our matching graph representation.

We evaluate our techniques on the publicly available Oxford buildings dataset (5K images) and show examples of objects automatically mined from this dataset. The methods are evaluated quantitatively on this dataset using a ground truth labelling for a number of Oxford landmarks. To demonstrate the scalability of the matching graph method, we show qualitative results on two larger datasets of images taken of the Statue of Liberty (37K images) and Rome (1M+ images).

6.2.2. Infinite Images: 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 for generating “infinite” images from large collections of photos by means of transformed image retrieval. Given a query image, we first transform it to simulate how it would look if the camera moved sideways and then perform image retrieval based on the transformed image. We then blend the query and retrieved images to create a larger panorama. Repeating this process will produce an “infinite" image. The transformed image retrieval model is not limited to simple 2D left/right image translation, however, and we show how to approximate other camera motions like rotation and forward motion/zoom-in using simple 2D image transforms. We represent images in the database as a graph where each node is an image and different types of edges correspond to different types of geometric transformations simulating different camera motions. Generating infinite images is thus reduced to following paths in the image graph. Given this data structure we can also generate a panorama that connects two query images, simply by finding the shortest path between the two in the image graph. We call this option the “image taxi". Our approach does not assume 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 synthesize new virtual scenes by combining images from distinct but visually similar locations. There are a number of potential applications to this technology. It can be used to generate long panoramas as well as content aware transitions between reference images or video shots. Finally, the image graph allows users to interactively explore large photo collections for ideation, games, social interaction and artistic purposes.

6.2.3. *Get out of my picture! Internet based inpainting (O. Whyte, J. Sivic and A. Zisserman)*

We present a method to replace a user specified target region of a photograph by using other photographs of the same scene downloaded from the Internet via viewpoint invariant image search. Each of the retrieved images is first geometrically then photometrically registered with the query photograph. Geometric registration is achieved using multiple homographies and photometric registration is performed using a global affine transformation on image intensities. Each retrieved image proposes a possible solution for the target region. In the final step we combine these proposals into a single visually consistent result, using a Markov random field optimisation to choose seams between proposals, followed by gradient domain fusion. We demonstrate removal of objects and people in challenging photographs of Oxford landmarks containing complex image structures. Example result is shown in figure 4.

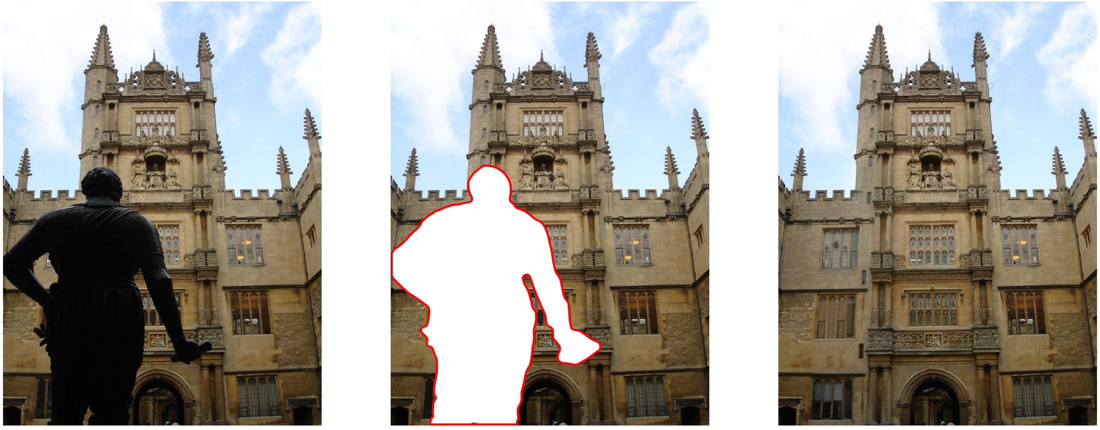


*Figure 4. **Automatic inpainting using photographs of the same scene downloaded from the Internet.** From left to right: Original query image, target region to replace, result from the system. The replaced region is consistent with the rest of the original image and the boundary between is effectively hidden, producing a convincing result. Note the complexity of image structures on the inpainted facade.*

6.2.4. *Avoiding confusing features in place recognition (J. Sivic, in collaboration with J. Knopp, CTU Prague / KU Lueven, and T. Pajdla, CTU Prague)*

We seek to recognize the place depicted in a query im- age using a database of Òstreet sideÓ images annotated with geolocation information. This is a challenging task due to changes in scale, viewpoint and lighting between the query and the images in the database. The image database may also contain objects, such as trees or road markings, which frequently occur and hence can cause significant confusion between different places. We employ the efficient bag-of- features representation previously used for object retrieval in large image collections. As the main contribution, we show how to avoid features leading to confusion of particular places by using geotags attached to database images as a form of supervision. We develop a method for automatic detection of image-specific and spatially-localized groups of confusing features, and demonstrate that suppressing them significantly improves place recognition performance while reducing the database size. As a second contribu- tion, we demonstrate that enhancing street side imagery with images downloaded from community photo-collections can lead to improved place recognition performance. Re- sults are shown on a geotagged database of over 17K im- ages of Paris downloaded from Google Street View.

6.2.5. Looking beyond image boundaries (J. Sivic, in collaboration with B. Kaneva, MIT, S. Avidan, Adobe, W. T. Freeman, MIT, and A. Torralba, MIT)

As we navigate through the world we constantly plan our next move based on what we would expect to see just around the corner or at the end of the hallway. Even in an unfamil- iar environment we still make predictions about what we might encounter using all of our prior experience. Here, we study a system to predict what is just beyond the image boundaries. The input to the system is a single image from a new environment and a large photo collection of images of the same class, but not of the exact same 3D location. The output is the image beyond the field of view of the query image according to different camera motions. To simulate a single motion - rotate or zoom-out - we first transform the input image, approximating what the camera would have seen under that motion, and then use the valid portion of the transformed image to perform the image retrieval. The key question we ask is: Can we predict what lies beyond image boundaries (the prediction problem)? As it turns out, this is related to our ability to match the transformed query image to images in the database (the matching problem). We quantify the quality of the system by comparing the re- trieval results to those of a ground truth data set that con- sists of geo-tagged images. This allows us to compare the predicted image to the actual image that was taken when the camera was actually moving. Our quantitative analysis provides an insight to what degree it is possible to predict image data beyond the image boundaries. We support our findings with a user study conducted using the Amazon Me- chanical Turk.

6.3. Learning image and object models

6.3.1. Non-local Sparse Models for Image Restoration (J. Mairal, F. Bach, J. Ponce, joint work with G. Sapiro, University of Minnesota)

In this work, we unify two different approaches to image restoration: On the one hand, learning a basis set (dictionary) adapted to sparse signal descriptions has proven to be very effective in image reconstruction and classification tasks. On the other hand, explicitly exploiting the self-similarities of natural images has led to the successful non-local means approach to image restoration. We pro- pose simultaneous sparse coding as a framework for combining these two approaches in a natural manner. This is achieved by jointly decomposing groups of similar signals on subsets of the learned dictionary. Experimental results in image denoising and demosaicking tasks with synthetic and real noise show that the proposed method outperforms the state of the art, making it possible to effectively restore raw images from digital cameras at a reasonable speed and memory cost.

6.3.2. Learning mid-level features for recognition (Y.-L. Boureau and J. Ponce, joint work with Y. LeCun, New York University)

Powerful handcrafted image descriptors developed in recent years have led to tremendous progress in recognition performance. They share common characteristics that can be described in a unified framework, as a three-stage pipeline of local feature extraction, pointwise non-linear transformation, and pooling of local features over some larger neighborhood. By contrast, the intermediate transformations that are then applied to the descriptors to form a suitable input for image classification are often cruder, i.e., variations of vector quantization. Using SIFT descriptors as input, we show in [24] that generalizing the three-stage pipeline to mid-level feature learning leads to state-of-the-art performance or better on several recognition benchmarks.

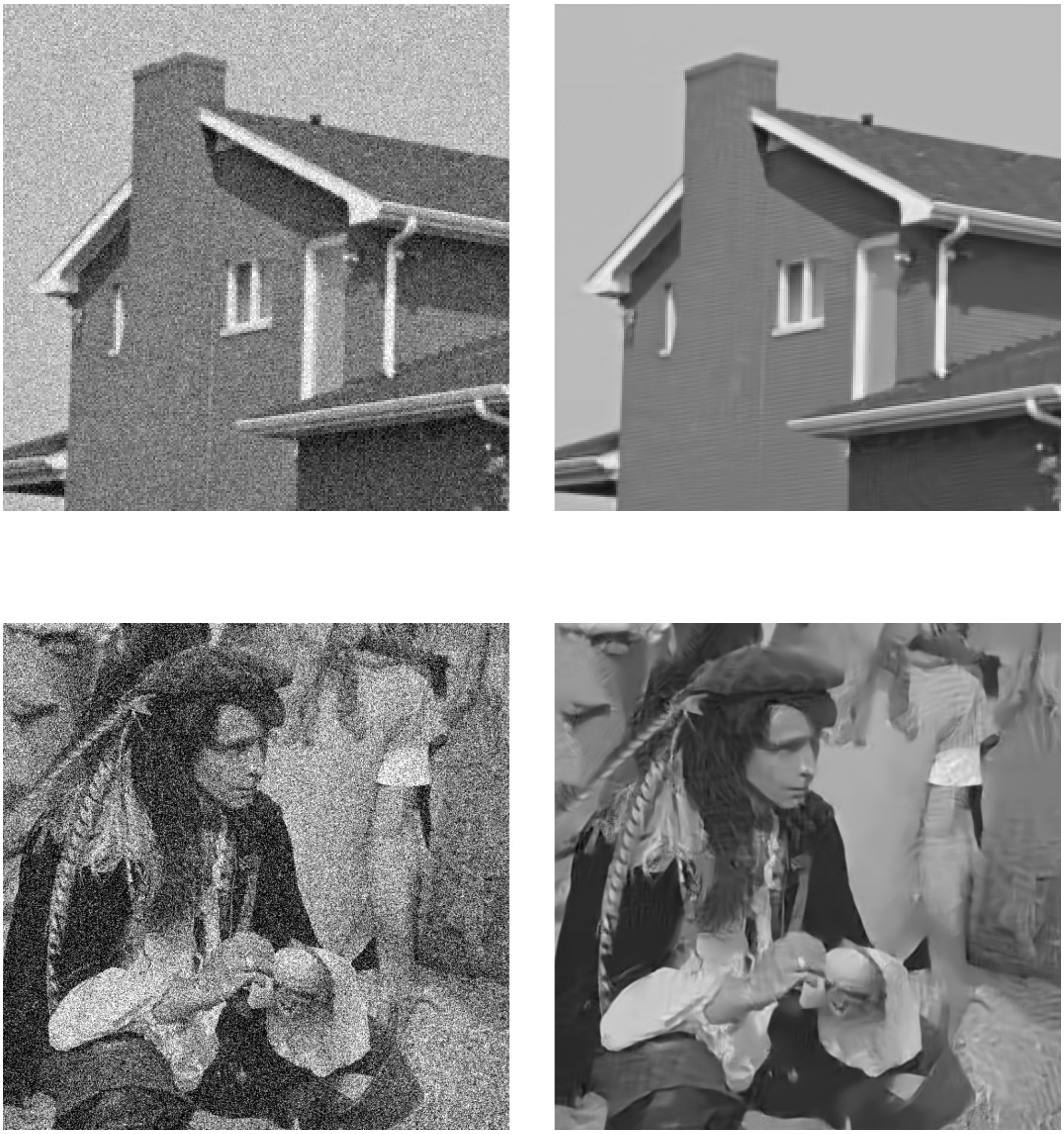


Figure 5. Qualitative evaluation of our denoising method with standard images. Left: noisy images. Right: restored images. Note that we reproduce the original brick texture in the house image ($\sigma = 15$) and the hair texture for the man image ($\sigma = 50$), both hardly visible in the noisy images. (The details are better seen by zooming on a computer screen.)

Performance increases when switching from hard to soft vector quantization, to unsupervised sparse coding, and finally to discriminative sparse coding. Moreover, we compare average and max pooling empirically and theoretically. Finally, we show that representing jointly small neighborhoods of SIFT by a single feature improves performance. Overall, lifting restrictions on the form of intermediate features to keep the same flexibility as when learning low-level image descriptors leads to high gains in recognition performance.

6.3.3. Reasoning About Object Relationships (A. Efros, with T. Malisiewicz, CMU)

The use of context is critical for scene understanding in computer vision, where the recognition of an object is driven by both local appearance and the object's relationship to other elements of the scene (context). Most current approaches rely on modeling the relationships between object categories as a source of context. In this paper we seek to move beyond categories to provide a richer appearance-based model of context. We present an exemplar-based model of objects and their relationships, the Visual Memex, that encodes both local appearance and 2D spatial context between object instances. We evaluate our model on Torralba's proposed Context Challenge against a baseline category-based system. Our experiments suggest that moving beyond categories for context modeling appears to be quite beneficial, and may be the critical missing ingredient in scene understanding systems.

6.3.4. Non-uniform Deblurring for Shaken Images (O. Whyte, J. Sivic, A. Zisserman and J. Ponce)

We argue that blur resulting from camera shake is mostly due to the 3D rotation of the camera, causing a blur that can be signiÞcantly non-uniform across the image. How- ever, most current deblurring methods model the observed image as a convolution of a sharp image with a uniform blur kernel. We propose a new parametrized geometric model of the blurring process in terms of the rotational velocity of the camera during exposure. We apply this model in the context of two different algorithms for camera shake removal: the Þrst uses a single blurry image (blind deblurring), while the second uses both a blurry image and a sharp but noisy im- age of the same scene. We show that our approach makes it possible to model and remove a wider class of blurs than previous approaches, and demonstrate its effectiveness with experiments on real images.

6.3.5. Segmenting Scenes by Matching Image Composites (B. Russell, A. Efros, J. Sivic and A. Zisserman, joint work with W. T. Freeman, MIT)

In this paper, we investigate how, given an image, similar images sharing the same global description can help with unsupervised scene segmentation. In contrast to recent work in semantic alignment of scenes, we allow an input image to be explained by partial matches of similar scenes. This allows for a better explanation of the input scenes. We perform MRF-based segmentation that optimizes over matches, while respecting boundary information. The recovered segments are then used to re-query a large database of images to retrieve better matches for the target regions. We show improved performance in detecting the principal occluding and contact boundaries for the scene over previous methods on data gathered from the LabelMe database.

6.3.6. Learning from Ambiguously Labeled Images (T. Cour, B. Sapp, C. Jordan and B. Taskar)

In many image and video collections, we have access only to partially labeled data. For example, personal photo collections often contain several faces per image and a caption that only specifies who is in the picture, but not which name matches which face. Similarly, movie screenplays can tell us who is in the scene, but not when and where they are on the screen, see figure 6. In [26], we formulate the learning problem in this setting as partially-supervised multiclass classification where each instance is labeled ambiguously with more than one label. We show theoretically that effective learning is possible under reasonable assumptions even when all the data is weakly labeled. We apply our framework to identifying faces culled from web news sources and to naming characters in TV series and movies, achieving 6% error for character naming on 16 episodes of LOST.

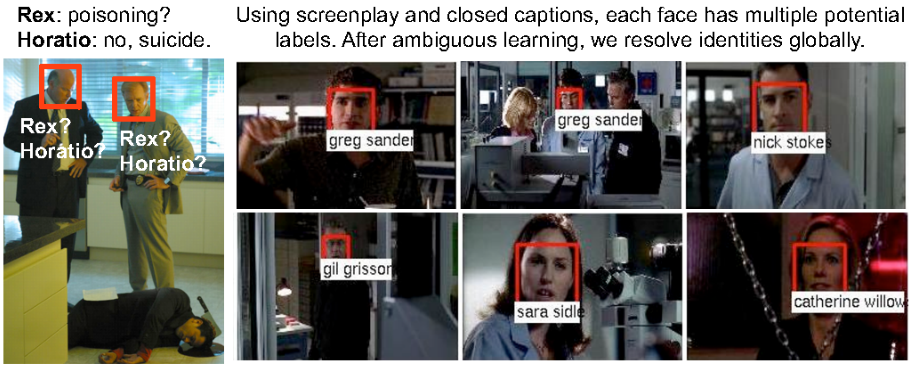

Rex: poisoning?
Horatio: no, suicide.
Using screenplay and closed captions, each face has multiple potential labels. After ambiguous learning, we resolve identities globally.
Rex?
Horatio?
Rex?
Horatio?
greg sande
greg sande
nick stoke
gil grissor
sara sidle
catherine willow

Figure 6. Learning from ambiguously labeled images

6.3.7. Learning discriminative part-based object models from weakly annotated images (T. Cour and F. Bach)

In many image recognition tasks, one key difficulty is the cost of annotating training examples with a bounding box or a segmentation. On the other hand, weakly annotated datasets composed of images and a set of attached labels are plentiful. Learning precise part-based object models from such weak supervision is very challenging, due to the lack of correspondences in the training set. We propose a discriminative learning of object parts based on a novel boosting formulation for multiple instance learning. We show our algorithm will produce a consistent labeling under certain restrictive but plausible conditions. We demonstrate our approach on a variety of weakly annotated image datasets.

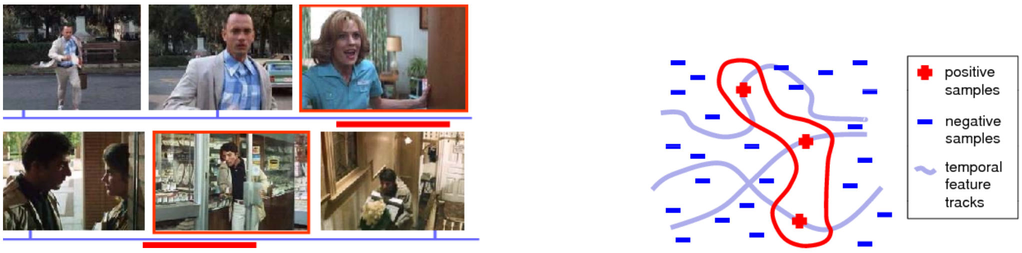
6.4. Human activity capture and classification

6.4.1. Automatic Annotation of Human Actions in Video (O. Duchenne, I. Laptev, J. Sivic, F. Bach and J. Ponce)

Our work in [29] addresses the problem of automatic temporal annotation of realistic human actions in video using minimal manual supervision. To this end we consider two associated problems: (a) weakly-supervised learning of action models from readily available annotations, and (b) temporal localization of human actions in test videos. To avoid the prohibitive cost of manual annotation for training, we use movie scripts as a means of weak supervision. Scripts, however, provide only implicit, sometimes noisy, and imprecise information about the type and location of actions in video (cf. Figure 7(a)). We address this problem with a kernel-based discriminative clustering algorithm that locates actions in the weakly-labeled training data (cf. Figure 7(b)). Using the obtained action samples, we train temporal action detectors and apply them to locate actions in the raw video data. Our experiments demonstrate that the proposed method for weakly-supervised learning of action models leads to significant improvement in action detection. We present detection results for three action classes in four feature length movies with challenging and realistic video data.

6.4.2. Actions in Context (I. Laptev, joint work with M. Marszałek and C. Schmid, INRIA-Grenoble)

We exploit the context of natural dynamic scenes for human action recognition in video. Human actions are frequently constrained by the purpose and the physical properties of scenes and demonstrate high correlation with particular scene classes. For example, eating often happens in a kitchen while running is more common outdoors (cf. Figure 8). The contribution of [40] is three-fold: (a) we automatically discover relevant scene



(a) (b)

Figure 7. (a): Video clips with OpenDoor actions provided by automatic script-based annotation. Selected frames illustrate both the variability of action samples within a class as well as the imprecise localization of actions in video clips. (b): In feature space, positive samples are constrained to be located on temporal feature tracks corresponding to consequent temporal windows in video clips. Background (non-action) samples provide further constrains on the clustering.

classes and their correlation with human actions, (b) we show how to learn selected scene classes from video without manual supervision and (c) we develop a joint framework for action and scene recognition and demonstrate improved recognition of both in natural video. We use movie scripts as a means of automatic supervision for training. For selected action classes we identify correlated scene classes in text and then retrieve video samples of actions and scenes for training using script-to-video alignment. Our visual models for scenes and actions are formulated within the bag-of-features framework and are combined in a joint scene-action SVM-based classifier. We report experimental results and validate the method on a new large dataset with twelve action classes and ten scene classes acquired from 69 movies.

(a) eating, kitchen (b) eating, cafe (c) running, road (d) running, street

Figure 8. Video samples from our dataset with high co-occurrences of actions and scenes and automatically assigned annotations.

6.4.3. *Evaluation of local spatio-temporal features for action recognition (I. Laptev, joint work with M. M. Ullah at INRIA-Rennes and H. Wang, A. Kläser C. Schmid at INRIA-Grenoble)*

Local space-time features have recently become a popular video representation for action recognition. Several methods for feature localization and description have been proposed in the literature and promising recognition results were demonstrated for a number of action classes. The comparison of existing methods, however, is often limited given the different experimental settings used. The purpose of this work [44] is to evaluate and compare previously proposed space-time features in a common experimental setup. In particular, we consider four different feature detectors and six local feature descriptors and use a standard bag-of-features SVM approach for action recognition. We investigate the performance of these methods on a total of 25 action classes distributed over three datasets with varying difficulty. Among interesting conclusions, we demonstrate that regular sampling of space-time features consistently outperforms all tested space-time interest point detectors for human actions in realistic settings. We also demonstrate a consistent ranking for the majority of methods over different datasets and discuss their advantages and limitations.

6.4.4. *Multi-view Synchronization of Human Actions and Dynamic Scenes (I. Laptev, joint work with E. Dexter and P. Pérez at INRIA-Rennes)*

This work deals with the temporal synchronization of image sequences. Two instances of this problem are considered: (a) synchronization of human actions and (b) synchronization of dynamic scenes with view changes. To address both tasks and to reliably handle large view variations, we in [27] use self-similarity matrices which remain stable across views. We propose time-adaptive descriptors that capture the structure of these matrices while being invariant to the impact of time warps between views. Synchronizing two sequences is then performed by aligning their temporal descriptors using the Dynamic Time Warping algorithm. We present quantitative comparison results between time-fixed and time-adaptive descriptors for image sequences with different frame rates. We also illustrate the performance of the approach on several challenging videos with large view variations, drastic independent camera motions and within-class variability of human actions.

6.4.5. *Quantitative analysis of videos for social sciences (N. Cherniavsky, I. Laptev, J. Ponce, J. Sivic, A. Zisserman)*

The display of human actions in mass media and its implications for our society is intensively studied in sociology, marketing and health care. For example, researchers have looked at the relationship between the incidence of characters who smoke in movies and adolescent smoking; the occurrence of drinking acts in movies and the consumption of alcohol; and the impact over time of the evolution of women activities depicted by TV shows. Video analysis for these purposes currently requires hours of tedious manual labeling, rendering large-scale experiments infeasible. Automating the detection and classification of human traits and actions in video will potentionally increase the quantity and diversity of experimental data. We are working with Institut National de l'Audiovisuel (INA), who has provided archive news footage for testing purposes, to automatically label people according to static and dynamic attributes, such as age, gender, race, clothing, hairstyle, and expression. It can be difficult to find enough good training data for such a specific milieu. We are exploring using transfer learning techniques to train a classifier from readily available still images from the web.

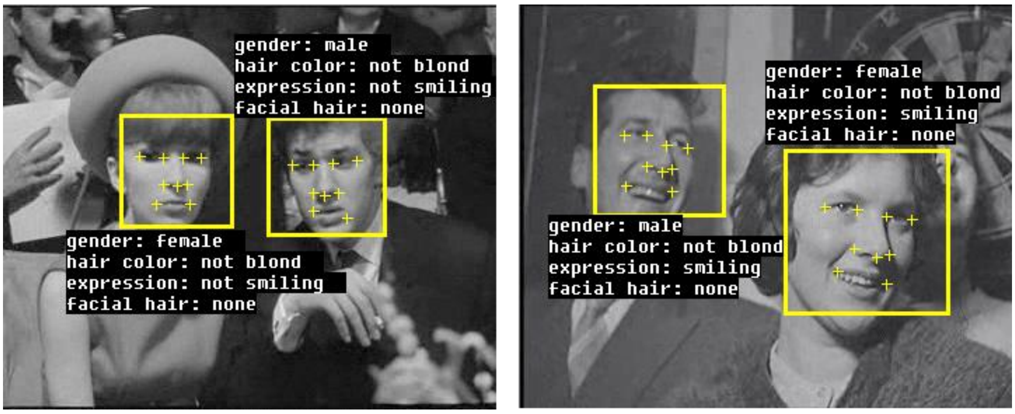


Figure 9. Sample frames from INA news footage videos with automatically detected faces and facial features overlaid. Text shows examples of considered attributes.

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

We investigate the problem of automatically labelling faces of characters in TV or movie material with their names, using only weak supervision from automatically-aligned subtitle and script text. Our previous work (Everingham et al.) 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.

In this paper we build 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, integration and recognition 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 seven episodes of the TV series “Buffy the Vampire Slayer”, demonstrating significantly increased coverage and performance with respect to previous methods on this material.

6.5. Machine learning for computer vision

6.5.1. *Online Matrix Factorization for Sparse Coding (J. Mairal, F. Bach, J. Ponce, joint work with G. Sapiro, University of Minnesota)*

Sparse coding—that is, modelling data vectors as sparse linear combinations of basis elements—is widely used in machine learning, neuroscience, signal processing, and statistics. This paper focuses on the large-scale matrix factorization problem that consists of learning the basis set, adapting it to specific data. Variations of this problem include dictionary learning in signal processing, non-negative matrix factorization and sparse principal component analysis. In this work, we propose to address these tasks with a new online optimization algorithm, based on stochastic approximations, which scales up gracefully to large datasets with millions of training samples, and extends naturally to various matrix factorization formulations, making it suitable for a wide range of learning problems. Experiments with natural images and genomic data demonstrates that it leads to state-of-the-art performance in terms of speed and optimization for both small and large datasets.

A software implementing these algorithms has been developed and registered at APP under the name SPAMS (Sparse Modeling Software).

6.5.2. *Vanishing point detection for road detection (H. Kong, J.-Y. Audibert and J. Ponce)*

Given a single image of an arbitrary road, that may not be well-paved, or have clearly delineated edges, or some a priori known color or texture distribution, is it possible for a computer to find this road? In [35], we address this question by decomposing the road detection process into two steps: the estimation of the vanishing point associated with the main (straight) part of the road, followed by the segmentation of the corresponding road area based on the detected vanishing point. The main technical contributions of the proposed approach are a novel adaptive soft voting scheme based on variable-sized voting region using confidence-weighted Gabor filters, which compute the dominant texture orientation at each pixel, and a new vanishing-point-constrained edge detection technique for detecting road boundaries. The proposed method has been implemented, and experiments with 1003 general road images demonstrate that it is both computationally efficient and effective at detecting road regions in challenging conditions (see Figure 11).

6.5.3. *Transductive segmentation of textured meshes (J.-Y. Audibert, joint work with A.-L. Jachiet, J.-P. Pons and R. Keriven)*

In [25], we address the problem of segmenting a textured mesh into objects or object classes, consistently with user-supplied seeds. We view this task as transductive learning and use the flexibility of kernel-based weights to incorporate a various number of diverse features. Our method combines a Laplacian graph regularizer that enforces spatial coherence in label propagation and an SVM classifier that ensures dissemination of the seeds characteristics. Our interactive framework allows to easily specify classes seeds with sketches drawn

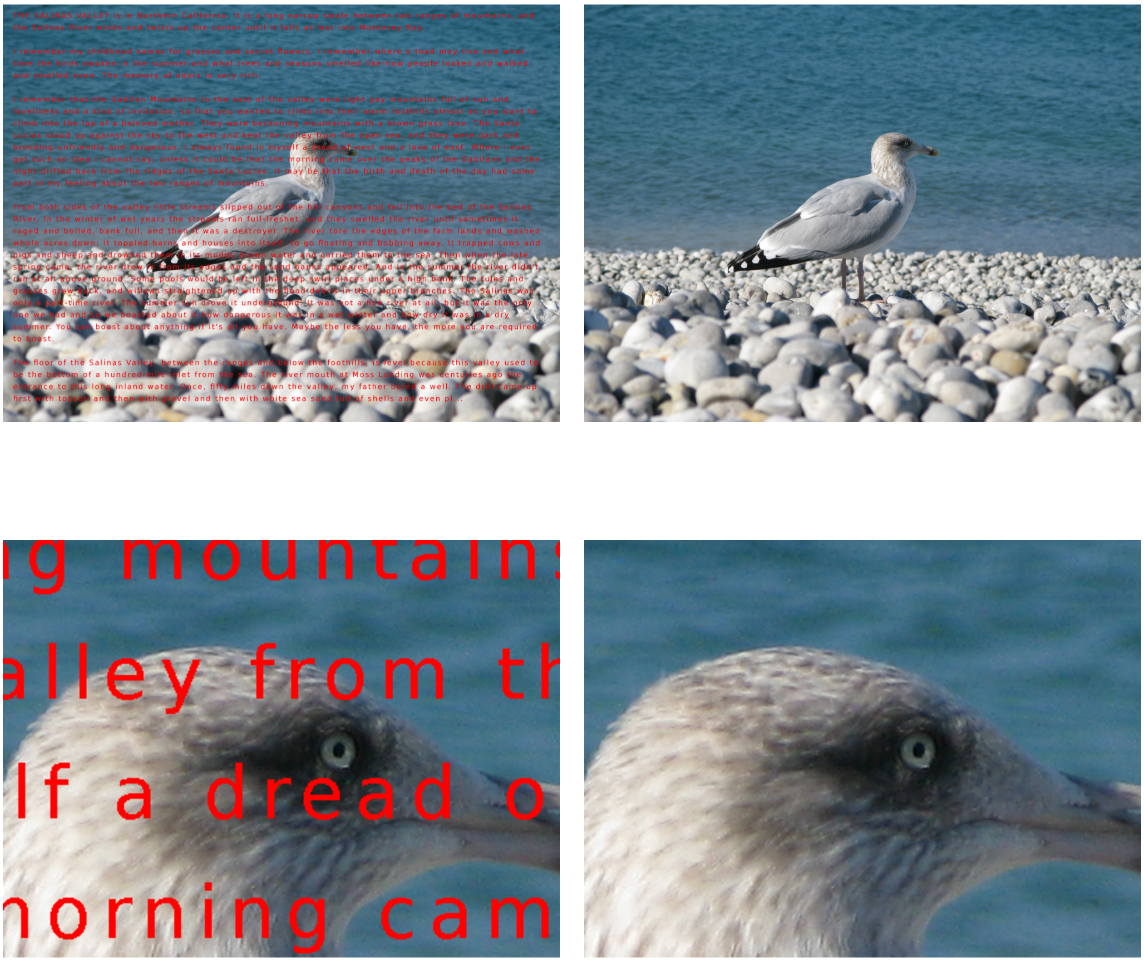


Figure 10. Inpainting example on a 12-Megapixel image using our fast online matrix factorization for sparse coding algorithms. Top: Damaged and restored images. Bottom: Zooming on the damaged and restored images. (Best seen in color).

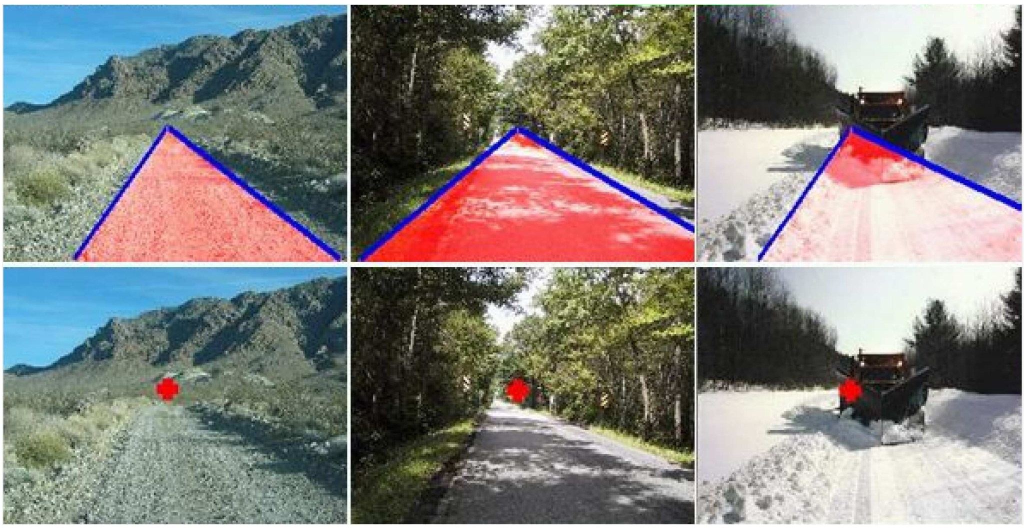


Figure 11. Vanishing point estimation and road detection

on the mesh and potentially refine the segmentation. We obtain qualitatively good segmentations on several architectural scenes and show the applicability of our method to outliers removing (see Figure 12).

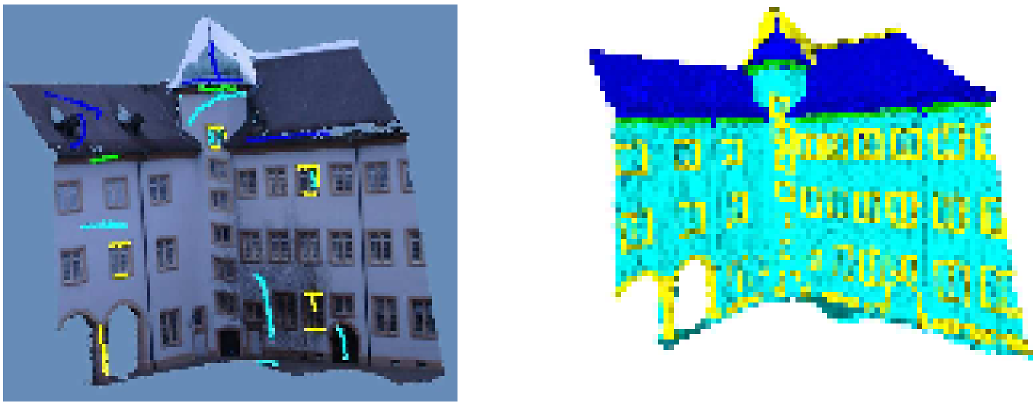


Figure 12. Segmentation of the mesh into four classes: roof, wall, windows edges, cornice. Left: the input textured mesh with user supplied sketches. Right: the resulting segmentation using our algorithm

6.6. Effective learning algorithms and architectures

6.6.1. Structured variable selection with sparsity-inducing norms (R. Jenatton, J-Y. Audibert and F. Bach)

We consider the empirical risk minimization problem for linear supervised learning, with regularization by structured sparsity-inducing norms. These are defined as sums of Euclidean norms on certain subsets of variables, extending the usual ℓ_1-norm and the group ℓ_1-norm by allowing the subsets to overlap. This leads to a specific set of allowed nonzero patterns for the solutions of such problems. We first explore the relationship between the groups defining the norm and the resulting nonzero patterns, providing both forward and backward algorithms to go back and forth from groups to patterns. This allows the design of norms adapted to specific prior knowledge expressed in terms of nonzero patterns. We also present an efficient active set algorithm, and analyze the consistency of variable selection for least-squares linear regression in low and high-dimensional settings.

6.6.2. Structured sparse principal component analysis (R. Jenatton, G. Obozinski and F. Bach)

We present an extension of sparse PCA, or sparse dictionary learning, where the sparsity patterns of all dictionary elements are structured and constrained to belong to a prespecified set of shapes. This *structured sparse PCA* is based on a structured regularization recently introduced by [50]. While classical sparse priors only deal with *cardinality*, the regularization we use encodes higher-order information about the data. We propose an efficient and simple optimization procedure to solve this problem. Experiments with two practical tasks, face recognition and the study of the dynamics of a protein complex, demonstrate the benefits of the proposed structured approach over unstructured approaches.

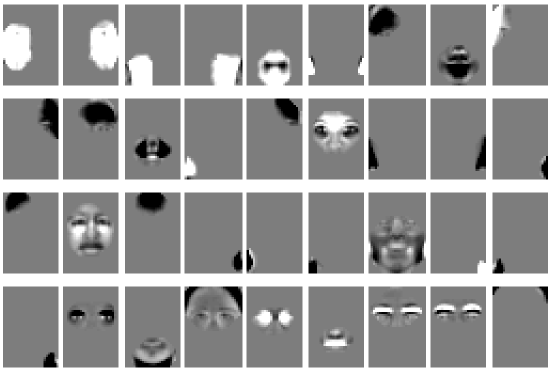


Figure 13. Exemple of a dictionary learned on a face database.

6.6.3. Group Lasso with Overlap and Graph Lasso (G. Obozinksi, joint work with L. Jacob and J.-P. Vert, Ecole des Mines de Paris

Modeling techniques that yield sparse models usually do not take into account available structural information on variables such as the fact that predictive features are often organised in groups or on graphs. For example, in computer vision, features associated to pixel lie naturally on a grid, discontinuities in the image often lie on curves, in computational biology predictive genes lie on regulation or interaction networks. In all these applications it is expected that relevant features are typically highly connected or concentrated in few groups. We developped in [31] regularization schemes that encode this prior information in the feature selection process to obtain sparse models which respect the structure and are therefore more accurate and more interpretable. We are considering the possibility of applying these models to contour extraction in images.

6.6.4. Global alignment of protein interaction networks by graph matching methods (F. Bach, in collaboration with M. Zaslavskiy and J.-P. Vert, Ecole des Mines de Paris)

Aligning protein-protein interaction (PPI) networks of different species has drawn a considerable interest recently. This problem is important to investigate evolutionary conserved pathways or protein complexes across species, and to help in the identification of functional orthologs through the detection of conserved interactions. It is however a difficult combinatorial problem, for which only heuristic methods have been proposed so far. We reformulate the PPI alignment as a graph matching problem, and investigate how state-of-the-art graph matching algorithms can be used for that purpose. We differentiate between two alignment problems, depending on whether strict constraints on protein matches are given, based on sequence similarity, or whether the goal is instead to find an optimal compromise between sequence similarity and interaction conservation in the alignment. We propose in [20] new methods for both cases, and assess their performance on the alignment of the yeast and fly PPI networks. The new methods consistently outperform state-of-the-art algorithms, retrieving in particular 78% more conserved interactions than IsoRank for a given level of sequence similarity. The source code for all conducted experiments is available at http://cbio.ensmp.fr/proj/graphm_ppi/.

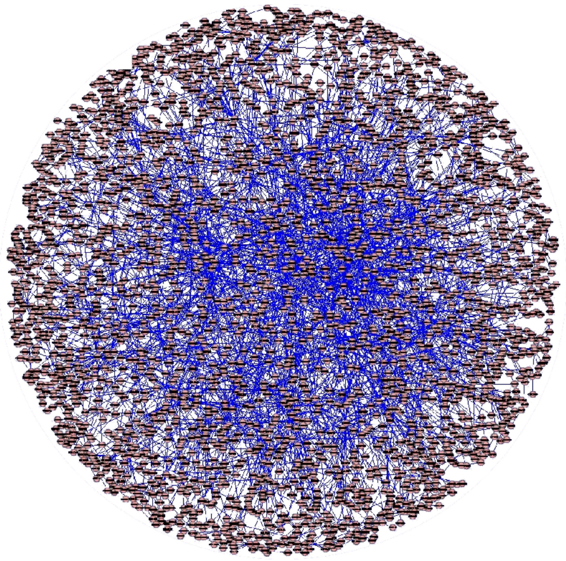


Figure 14. Fly protein-protein interaction network. Vertices (nodes) represent proteins and edges correspond to protein-protein interactions.

6.7. Learning theory

6.7.1. Data-driven calibration of linear estimators with minimal penalties (S. Arlot and F. Bach)

We tackle the problem of selecting among several linear estimators in non-parametric regression; this includes model selection for linear regression, the choice of a regularization parameter in kernel ridge regression or spline smoothing, and the choice of a kernel in multiple kernel learning. We propose a new algorithm which first estimates consistently the variance of the noise, based upon the concept of minimal penalty which was previously introduced in the context of model selection. Then, plugging our variance estimate in Mallows' C_L penalty is proved to lead to an algorithm satisfying an oracle inequality. Simulation experiments with kernel ridge regression and multiple kernel learning show that the proposed algorithm often improves significantly existing calibration procedures such as 10-fold cross-validation or generalized cross-validation [21].

6.7.2. Resampling-based estimation of the accuracy of satellite ephemerides (S. Arlot, joint work with J. Desmars, J.-E. Arlot, V. Lainey and A. Vienne)

The accuracy of predicted orbital positions depends on the quality of the theorical model and of the observations used to fit the model. During the period of observations, this accuracy can be estimated through comparison with observations. Outside this period, the estimation remains difficult. Many methods have been developed for asteroid ephemerides in order to evaluate this accuracy. We introduced in [5] a new method for estimating the accuracy of predicted positions at any time, in particular outside the observation period. This new method is based upon a bootstrap resampling and allows this estimation with minimal assumptions. The method was applied to two of the main Saturnian satellites, Mimas and Titan, and compared with other methods used previously for asteroids. The bootstrap resampling is a robust and practical method for estimating the accuracy of predicted positions.

6.7.3. Asymptotically optimal regularization in smooth parametric models. (F. Bach, with P. Liang, G. Bouchard, M. I. Jordan)

Many types of regularization schemes have been employed in statistical learning, each one motivated by some assumption about the problem domain. In [36], we present a unified asymptotic analysis of smooth regularizers, which allows us to see how the validity of these assumptions impacts the success of a particular regularizer. In addition, our analysis motivates an algorithm for optimizing regularization parameters, which in turn can be analyzed within our framework. We apply our analysis to several examples, including hybrid generative-discriminative learning and multi-task learning.

6.7.4. Minimax policies for adversarial and stochastic bandits (J.-Y. Audibert)

In the multi-armed bandit problem, at each stage, an agent (or decision maker) chooses one action (or arm), and receives a reward from it. The agent aims at maximizing his rewards. Since he does not know the process generating the rewards, he needs to explore (try) the different actions and yet, exploit (concentrate its draws on) the seemingly most rewarding arms. In [22], we fill in a long open gap in the characterization of the minimax rate for the multi-armed bandit problem. Concretely, we remove an extraneous logarithmic factor in the previously known upper bound and propose a new family of randomized algorithms based on an implicit normalization, as well as a new analysis. We also consider the stochastic case, and prove that an appropriate modification of the upper confidence bound policy UCB1 (Auer et al., 2002) achieves the distribution-free optimal rate while still having a distribution-dependent rate logarithmic in the number of plays.

6.7.5. Linear least squares regression (J.-Y. Audibert)

In [49], we consider the problem of predicting as well as the best linear combination of d given functions in least squares regression, and variants of this problem including constraints on the parameters of the linear combination. When the input distribution is known, there already exists an algorithm having an expected excess risk of order d/n, where n is the size of the training data. Without this strong assumption, standard results often contain a multiplicative $\log n$ factor, and require some additional assumptions like uniform boundedness of the d-dimensional input representation and exponential moments of the output. This work provides new risk bounds for the ridge estimator and the ordinary least squares estimator, and their variants. It also provides shrinkage procedures with convergence rate d/n (i.e., without the logarithmic factor) in expectation and in deviations, under various assumptions. The key common surprising factor of these results is the absence of exponential moment condition on the output distribution while achieving exponential deviations. All risk bounds are obtained through a PAC-Bayesian analysis on truncated differences of losses. Finally, we show that some of these results are not particular to the least squares loss, but can be generalized to similar strongly convex loss functions.

6.7.6. Change-point detection (S. Arlot, joint work with A. Celisse)

We tackle the problem of detecting abrupt changes in the mean of a heteroscedastic signal by model selection, without knowledge on the variations of the noise. A new family of change-point detection procedures is proposed, showing that cross-validation methods can be successful in the heteroscedastic framework, whereas most existing procedures are not robust to heteroscedasticity. The robustness to heteroscedasticity of the proposed procedures is supported by an extensive simulation study, together with recent theoretical results. An application to Comparative Genomic Hybridization (CGH) data is provided, showing that robustness to heteroscedasticity can indeed be required for their analysis [48].

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 collaboration efforts via tutorial presentations and discussions with A. Zisserman, J. Sivic and J. Ponce at EADS and ENS. The tutorial was delivered at EADS Suresnes lab in January 2009 and covered face recognition in videos. In addition, Marc Sturzel (EADS) is doing a PhD at ENS with Jean Ponce and Andrew Zisserman.

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

Participants: Jean Ponce, Francis Bach, Andrew Zisserman, Josef Sivic, Ivan Laptev.

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, Oliver 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, in Willow, we focus on high-fidelity image-based modeling, 3D shape/appearance matching and image/video enhancement and restoration from multiple un-calibrated photographs. The goal is to 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: DETECT (ENS)

Participants: Sylvain Arlot, Francis Bach, Josef Sivic.

The DETECT project aims at providing new statistical approaches for detection problems in computer vision (in particular, detecting and recognizing human actions in videos) and bioinformatics (e.g., simultaneously segmenting CGH profiles). These problems are mainly of two different statistical nature: multiple change-point detection (i.e., partitioning a sequence of observations into homogeneous contiguous segments) and multiple tests (i.e., controlling a priori the number of false positives among a large number of tests run simultaneously).

This is a collaborative effort with A. Celisse (University Lille 1), T. Mary-Huard (AgroParisTech), E. Roquain and F. Villers (Univeristy Paris 6), in addition to S. Arlot, F. Bach and J. Sivic from Willow.

S. Arlot is the leader of this ANR "Young researchers" project. The total funding is 70000 Euros.

8.3. 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).

8.4. 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.

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

Participant: Francis Bach.

This is 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.

8.6. European Research Council (ERC) Starting Investigator Researcher grant

Participant: Francis Bach.

SIERRA is a research project funded by the European Research Council (ERC) and coordinated by Francis Bach. It is located within the joint INRIA/CNRS/Ecole Normale Superieure computer science laboratory in downtown Paris. The goals of the project are to explore sparse structured methods for machine learning, with applications in computer vision and audio processing.

9. Dissemination

9.1. Leadership within the scientific community

- Conference and workshop organization:
  - Co-organizer, Conference on Learning Theory Workshop on On-line Learning with Limited Feedback, Montreal, Canada, 2009, (J.-Y. Audibert)
  - Co-organizer, Pascal VOC 2009 Workshop at the IEEE International Conference on Computer Vision, Kyoto, Japan, 2009 http://pascallin.ecs.soton.ac.uk/challenges/VOC/ (A. Zisserman).
  - Co-organizer, International Workshop on Video, Barcelona, Spain, 2009 http://research.microsoft.com/en-us/um/india/events/iwv09/index.html (A. Zisserman).
  - Co-organizer, NIPS 2009 Workshop on “Understanding multiple kernel learning methods”:http://mkl.ucsd.edu/workshop (F. Bach).
- Editorial boards:
  - Journal of Machine Learning Research, Action Editor (F. Bach).
  - International Journal of Computer Vision (J. Ponce, A. Zisserman).
  - Fondations and Trends in Computer Graphics and Vision (J. Ponce).
  - SIAM Journal on Imaging Sciences (J. Ponce, F. Bach)
  - Action editor, Journal of Machine Learning Research (F. Bach)
  - Associate editor, IEEE Transactions Pattern Analysis and Machine Intelligence (F. Bach)
  - Associate Editor of Image and Vision Computing Journal (I. Laptev)
- Area chairs:
  - Neural Information and Processing Systems (NIPS) Conference, 2009 (J.-Y. Audibert, A. Zisserman)
  - IEEE Conference on Computer Vision and Pattern Recognition, 2010 (I. Laptev).
  - International Conference on Computer Vision, 2009 (J. Ponce).
  - International Conference on Machine Learning, 2009 (F. Bach)
  - European Conference on Machine Learning, 2009 (F. Bach)
- Program committees:

  - Conférence francophone sur l'apprentissage automatique (CAP), 2009 (J.-Y. Audibert)
  - IEEE Conference on Computer Vision and Pattern Recognition, 2009 (J. Sivic, F. Bach, T. Cour, I. Laptev).
  - IEEE International Conference on Computer Vision, 2009 (J. Sivic, F. Bach, T. Cour, I. Laptev).
  - ACM International Conference on Image and VIdeo Retrieval, 2009 (J. Sivic).
  - Reviewer for the Neural Information and Processing Systems (NIPS), 2009 (J. Sivic, T. Cour, J. Mairal).
  - Reviewer for the ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), 2009 (J. Sivic).
  - British Machine Vision Conference, 2009 (I. Laptev).

- Prizes:
  - The paper *Minimax policies for adversarial and stochastic bandits* by J.-Y. Audibert and S. Bubeck was awarded the best student paper at the Conference on Learning Theory (COLT), 2009
  - The paper *A tensor-based algorithm for high-order graph matching* by O. Duchenne, F. Bach, I. Kweon, and J. Ponce, has been awarded a Best Student Paper award (Honorable Mention).
  - Andrew Zisserman received the BMVA Distinguished Fellow award, 2009.

- 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 and J.-Y. Audibert co-coordinate the Paris reading group/seminar in machine learning (http://sites.google.com/site/smileinparis/).
  - 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 serves on the contents commission of Cap Digital.
  - 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

- S. Arlot (together with G. Stoltz), "Supervised classification: algorithms and their data-driven calibration", École Centrale de Paris, 3rd year, 16h.
- J.-Y. Audibert, "Machine Learning and applications", Ecole des Ponts ParisTech, 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.
- F. Bach, NIPS 2009 Tutorial on sparse methods for machine learning
- F. Bach, J. Mairal, J. Ponce, ICCV 2009 Tutorial on Sparse Coding and Dictionary Learning for Image Analysis
- 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.
- I. Laptev, 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, 33h.
- A. Zisserman, Lecture course on “Computer Vision”, University of Oxford.
- A. Zisserman, Lecture course on “Machine Learning”, University of Oxford.
- A. Zisserman, Lecture course on “Optimization”, University of Oxford.

9.3. Invited presentations

- S. Arlot, *Data-driven penalties for model selection*, Mathematical Statistics Seminar, Weierstrass Institute for Applied Analysis and Stochastics, Berlin, April 2009.
- S. Arlot, *Optimal model selection*, European Meeting of Statisticians, Toulouse, July 2009.
- J.-Y. Audibert, *Bornes sur le risque en régression linéaire*, Groupe de Travail de l’ENSAE, Paris, May 2008
- J.-Y. Audibert, *PAC-Bayes bounds*, University College London, Great Britain, Jul. 2009
- J.-Y. Audibert, *Risk bounds in linear regression through PAC-Bayesian truncation*, Séminaire du LSTA, Paris, Oct. 2009
- F. Bach, Workshop “Sparsity in Machine Learning and Statistics”, Cumberland Lodge, UK, 2009
- F. Bach, ICML workshop on feature hierarchies, Montréal, Canada
- F. Bach, Journées Françaises de Statistique, Bordeaux, 2009
- F. Bach, Séminaire de statistique, Université Paul Sabatier, Toulouse, 2009
- J. Mairal, Sparse learned representations for image restoration. Symposium Patch-based Image Representation, Manifolds and Sparsity, 2009. Rennes.
- T. Cour, CVPR 2009 Workshop on Visual and Contextual Learning from Annotated Images and Videos.
- J. Ponce, Laboratoire d’Analyse et d’Architecture des Systèmes
- J. Ponce, Télécom Paristech
- J. Ponce, Beckman Institute 20th Anniversary Symposium
- J. Ponce, ICCV Area Chair Symposium
- J. Ponce, University of Southern California
- J. Sivic, INRIA Rennes seminar, S. Malo, France.
- J. Sivic, INRIA Grenoble seminar, Grenoble, France.
- J. Sivic, International workshop on video, Barcelona.
- J. Sivic, BIRS Workshop on Computer Vision and the Internet, Banff, Canada.
- L. Benoit, N. Cherniavsky, J. Sivic, Salon europeen de la Recherche et de l’Innovation, Paris
- B. Russell, invited speaker at the One day workshop, Czech Technical University in Prague

- B. Russell, invited speaker at the ICCV Workshop on 3D Representation for Recognition, Kyoto, Japan
- I. Laptev, Keynote at Sibgrapi 2009 Digital Video Journey, Rio de Janeiro, Brazil, October 2009
- I. Laptev, Workshop on Trends in Computer Vision, Prague, July 2009
- I. Laptev, International Workshop on Video, Barcelona, May 2009
- A. Zisserman, Key note speaker at the International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), London.
- A. Zisserman, Invited presentation at BIRS Workshop on Computer Vision and the Internet, Banff, Canada.

9.4. Popular science

- J. Ponce et L. Benoit, intervention in Emission spéciale en public et en direct de l'ENS Ulm : "Cryptographie et vision artificielle", Place de la Toile, France Culture, Sep. 25, 2009.
- J. Ponce, "Comment donner un sens á l'image numérique?", Les cahiers de l'INRIA, La Recherche, Nov. 2009, No. 435.

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", vol. 10, 2009, 803—826.

[2] S. ARLOT. *Model selection by resampling penalization*, in "Electron. J. Statist.", vol. 3, 2009, p. 557–624 (electronic), http://www.di.ens.fr/willow/pdfs/2009_Arlot_EJS-2008-196.pdf.

[3] S. ARLOT, G. BLANCHARD, É. ROQUAIN. *Some non-asymptotic results on resampling in high dimension, I: confidence regions*, in "Ann. Statist.", 2009, To appear.

[4] S. ARLOT, G. BLANCHARD, É. ROQUAIN. *Some non-asymptotic results on resampling in high dimension, II: multiple tests*, in "Ann. Statist.", 2009, To appear.

[5] J. DESMARS, S. ARLOT, J.-E. ARLOT, V. LAINEY, A. VIENNE. *Estimating the accuracy of satellite ephemerides using the bootstrap method*, in "Astronomy and Astrophysics", vol. 499, May 2009, p. 321–330, http://www.di.ens.fr/willow/pdfs/2009_AandA_Desmars_Arlot_etal.pdf.

[6] M. EVERINGHAM, J. SIVIC, A. ZISSERMAN. *Taking the bite out of automatic naming of characters in TV video*, in "Image and Vision Computing", vol. 27, n^{o} 5, 2009, p. 545–559.

[7] K. FUKUMIZU, F. BACH, M. I. JORDAN. *Kernel dimension reduction in regression*, in "Annals of Statistics", vol. 37, n^{o} 4, 2009, 1871—1905.

[8] Y. FURUKAWA, J. PONCE. *Accurate Camera Calibration from Multi-View Stereo and Bundle Adjustment*, in "International Journal of Computer Vision", vol. 84, n^{o} 3, 2009, p. 257-268.

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

[10] Y. FURUKAWA, J. PONCE. *Accurate, Dense, and Robust Multi-View Stereopsis*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2009, In press..

[11] Y. FURUKAWA, J. PONCE. *Carved Visual Hulls for Image-Based Modeling*, in "International Journal of Computer Vision", vol. 81, n^{o} 1, 2009, p. 53-67.

[12] B. KANEVA, J. SIVIC, A. TORRALBA, S. AVIDAN, W. FREEMAN. *InÞnite Images: Creating and Exploring a Large Photorealistic Virtual Space*, in "Proceedings of the IEEE", 2009, Accepted..

[13] H. KONG, J.-Y. AUDIBERT, J. PONCE. *General road detection from a single image*, in "IEEE Transactions on Image Processing", 2009, Accepted pending minor revisions.

[14] J.-F. LALONDE, A. A. EFROS, S. G. NARASIMHAN. *Webcam Clip Art: Appearance and Illuminant Transfer from Time-lapse Sequences*, in "ACM Transactions on Graphics (SIGGRAPH Asia 2009)", vol. 28, n^{o} 5, December 2009.

[15] J. MAIRAL, F. BACH, J. PONCE, G. SAPIRO. *Online Learning for Matrix Factorization and Sparse Coding*, in "Journal of Machine Learning Research", 2009.

[16] J. PHILBIN, J. SIVIC, A. ZISSERMAN. *Geometric Latent Dirichlet Allocation on a Matching Graph for Large-Scale Image Datasets*, in "Int. J. of Comp. Vision", 2009, Submitted..

[17] J. PONCE. *Comment donner un sens à l'image numérique*, in "La Recherche – Les cahiers de l'INRIA", vol. 435, 2009.

[18] J. SIVIC, A. ZISSERMAN. *Efficient Visual Search Cast as Text Retrieval*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", vol. 31, n^{o} 4, 2009, p. 591–606.

[19] M. ZASLAVSKIY, F. BACH, J.-P. VERT. *A path following algorithm for the graph matching problem*, in "IEEE Trans. Patt. Anal. Mach. Intell.", vol. 31, n^{o} 12, 2009, 2227—2242.

[20] M. ZASLAVSKIY, F. BACH, J.-P. VERT. *Global alignment of protein-protein interaction networks by graph matching methods*, in "Bioinformatics", vol. 25, n^{o} 12, 2009, 1259—1267.

International Peer-Reviewed Conference/Proceedings

[21] S. ARLOT, F. BACH. *Data-driven calibration of linear estimators with minimal penalties*, in "Advances in Neural Information Processing Systems (NIPS)", December 2009.

[22] J.-Y. AUDIBERT, S. BUBECK. *Minimax policies for adversarial and stochastic bandits*, in "22th annual conference on learning theory, Montreal, Canada", Jun 2009.

[23] G. BATOG, X. GOAOC, J. PONCE. *Admissible Linear Map Models of Linear Cameras*, in "IEEE Conference on Computer Vision and Pattern Recognition", 2010, Submitted..

[24] Y.-L. BOUREAU, Y. LECUN, J. PONCE. *Learning Mid-Level Features for Recognition*, in "IEEE Conference on Computer Vision and Pattern Recognition", 2010, Submitted..

[25] A.-L. CHAUVE, J.-P. PONS, J.-Y. AUDIBERT, R. KERIVEN. *Transductive segmentation of textured meshes*, in "Asian Conference on Computer Vision, Xi' an, China", Sep 2009, http://imagine.enpc.fr/publications/papers/ACCV09a.pdf.

[26] T. COUR, B. SAPP, C. JORDAN, B. TASKAR. *Learning from Ambiguously Labeled Images*, in "IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'09)", 2009.

[27] E. DEXTER, P. PÉREZ, I. LAPTEV. *Multi-View Synchronization of Human Actions and Dynamic Scenes*, in "British Machine Vision Conference", 2009.

[28] O. DUCHENNE, F. BACH, I. KWEON, J. PONCE. *A Tensor-Based Algorithm for High-Order Graph Matching*, in "IEEE Conference on Computer Vision and Pattern Recognition", 2009.

[29] O. DUCHENNE, I. LAPTEV, J. SIVIC, F. BACH, J. PONCE. *Automatic Annotation of Human Actions in Video*, in "Proc. Int. Conf. Comp. Vision", 2009.

[30] Y. FURUKAWA, J. PONCE. *Dense 3D Motion Capture for Human Faces*, in "IEEE Conference on Computer Vision and Pattern Recognition", 2009.

[31] L. JACOB, G. OBOZINSKI, J.-P. VERT. *Group Lasso with Overlap and Graph Lasso*, in "Proceedings of the 26th Annual International Conference on Machine Learning (ICML)", 2009.

[32] A. JOULIN, F. BACH, J. PONCE. *Discriminative Clustering for Image Co-segmentation*, in "IEEE Conference on Computer Vision and Pattern Recognition", 2010, Submitted..

[33] B. KANEVA, S. AVIDAN, W. FREEMAN, J. SIVIC, A. TORRALBA. *Looking Beyond Image Boundaries*, in "Proc. IEEE Conf. Comp. Vision Patt. Recog.", 2010, Submitted..

[34] J. KNOPP, J. SIVIC, T. PAJDLA. *Avoiding confusing features in place recognition*, in "Proc. IEEE Conf. Comp. Vision Patt. Recog.", 2010, Submitted..

[35] H. KONG, J.-Y. AUDIBERT, J. PONCE. *Vanishing point detection for road detection*, in "Conference on Computer Vision and Pattern Recognition (CVPR), Miami, United States", Jun 2009.

[36] P. LIANG, F. BACH, G. BOUCHARD, M. I. JORDAN. *Asymptotically Optimal Regularization in Smooth Parametric Models*, in "Advances in Neural Information Processing Systems (NIPS)", 2009.

[37] J. MAIRAL, F. BACH, J. PONCE, G. SAPIRO. *Online Dictionary Learning for Sparse Coding*, in "Proc. Int. Conf. on Machine Learning", 2009.

[38] J. MAIRAL, F. BACH, J. PONCE, G. SAPIRO, A. ZISSERMAN. *Non-local Sparse Models for Image Restoration*, in "Proc. Int. Conf. Comp. Vision", 2009.

[39] T. MALISIEWICZ, A. A. EFROS. *Beyond Categories: The Visual Memex Model for Reasoning About Object Relationships*, in "NIPS", December 2009.

[40] M. MARSZAŁEK, I. LAPTEV, C. SCHMID. *Actions in Context*, in "Proc. IEEE Conf. Comp. Vision Patt. Recog.", 2009.

[41] J. PONCE. *What is a camera?*, in "IEEE Conference on Computer Vision and Pattern Recognition", 2009.

[42] B. RUSSELL, A. A. EFROS, J. SIVIC, W. FREEMAN, A. ZISSERMAN. *Segmenting Scenes by Matching Image Composites*, in "Proc. Neural Info. Proc. Systems", 2009.

[43] 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", 2009, submitted.

[44] H. WANG, M. ULLAH, A. KLÄSER, I. LAPTEV, C. SCHMID. *Evaluation of local spatio-temporal features for action recognition*, in "British Machine Vision Conference", 2009.

[45] O. WHYTE, J. SIVIC, A. ZISSERMAN, J. PONCE. *Non-Uniform Deblurring for Shaken Images*, in "IEEE Conference on Computer Vision and Pattern Recognition", 2010, Submitted..

[46] O. WHYTE, J. SIVIC, A. ZISSERMAN. *Get out of my picture! Internet-based inpainting*, in "British Machine Vision Conference", 2009.

Scientific Books (or Scientific Book chapters)

[47] S. LAZEBNIK, C. SCHMID, J. PONCE. *Spatial Pyramid Matching*, in "Object Categorization: Computer and Human Vision Perspectives", S. DICKINSON, A. LEORNADIS, B. SCHIELE, M. TARR (editors), Cambridge University Press, 2009.

Research Reports

[48] S. ARLOT, A. CELISSE. *Segmentation in the mean of heteroscedastic data via cross-validation*, ArXiv, April 2009, http://arxiv.org/pdf/0902.3977v2, arXiv:0902.3977v2, Technical report.

[49] J.-Y. AUDIBERT, O. CATONI. *Risk bounds in linear regression through PAC-Bayesian truncation*, Feb 2009, http://hal.archives-ouvertes.fr/docs/00/36/02/68/PDF/dovern.pdf, Technical report.

[50] R. JENATTON, J.-Y. AUDIBERT, F. BACH. *Structured Variable Selection with Sparsity-Inducing Norms*, arXiv:0904.3523, 2009, Technical report.

[51] R. JENATTON, G. OBOZINSKI, F. BACH. *Structured sparse principal component analysis*, arXiv:0909.1440, 2009, Technical report.