IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2011

Project-Team WILLOW

Models of visual object recognition and scene understanding

IN COLLABORATION WITH: Laboratoire d'Informatique de l'Ecole Normale Supérieure (LIENS)

RESEARCH CENTER
Paris - Rocquencourt

THEME
Vision, Perception and Multimedia Understanding

Table of contents

Project-Team WILLOW

Keywords: 3D Modeling, Classification, Computer Vision, Machine Learning, Recognition, Interpretation

1. Members

Research Scientists

Ivan Laptev [Chargé de Recherches INRIA]
Jean Ponce [Team Leader, Professor in the Département d'Informatique of École Normale Supérieure (ENS) [Habilite]
Josef Sivic [Chargé de Recherches INRIA]
Andrew Zisserman [Team Co-leader, Professor in the Engineering Department of the University of Oxford, and part-time professor at ENS, HdR]

PhD Students

Mathieu Aubry
Louise Benoît
Y-Lan Boureau
Florent Couzinié-Devy
Vincent Delaitre
Olivier Duchenne
Warith Harchaoui
Armand Joulin
Guillaume Seguin
Marc Sturzel
Muhammad Ullah
Oliver Whyte

Post-Doctoral Fellows

Karteek Alahari
Yves Ubelmann

Administrative Assistant

Marine Meyer

2. Overall Objectives

2.1. Statement

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

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

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

WILLOW was created in 2007: It was recognized as an INRIA team in January 2007, and as an official project-team in June 2007. WILLOW is a joint research team between INRIA Paris Rocquencourt, Ecole Normale Supérieure (ENS) and Centre National de la Recherche Scientifique (CNRS).

This year we have hired two new Phd students: Guillaume Seguin (ENS) and Mathieu Aubry (ENPC). Alexei Efros (Professor, Carnegie Mellon University, USA), Abhinav Gupta (Assistant Research Professor, Carnegie Mellon University, USA) visited WILLOW in summer 2011 together with their student Carl Doersch (CMU).

2.2. Highlights

+ Julien Mairal, a former PhD student of J. Ponce and F. Bach won several prizes for his PhD thesis about "Sparse coding for machine learning, image processing and computer vision". See details in section 9.1.
+ Jean Ponce was awarded an Advanced ERC Grant, starting Jan 2011.
+ Andrew Zisserman was awarded the Rank Prize for his "Outstanding contributions to modern computer vision" http://www.rankprize.org/.
+ I. Laptev, J. Ponce and J. Sivic (together with C. Schmid (INRIA Grenoble)) co-organized one week summer school on visual recognition and machine learning at Ecole Normale Superieure http://www.di.ens.fr/willow/events/cvml2011/. The school has attracted 175 participants from 28 countries.
+ The updated 2nd edition of the textbook "Computer Vision: A Modern Approach" by David Forsyth and Jean Ponce has been published by Pearson Education in November 2011.
+ The group has split into two on January 1st 2011 to create a new INRIA project-team called SIERRA. The new group and its interactions with WILLOW is described in section 6.5.

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), retrieval of particular objects and buildings from images and videos (Sivic and Zisserman, 2003) and (Philbin *et al.*, 2007), and a theoretical study of a general formalism for modeling central and non-central cameras using the formalism and terminology of classical projective geometry (Ponce, 2009 and Batog *et al.*, 2010).

We have also 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, outlined in detail in section 6.1, has focused on (i) using our multi-view-stereo approach to model archaeological sites together with developing representations and efficient retrieval techniques to enable matching historical paintings to 3D models of archeological sites, and (ii) visual place recognition in structured databases, where images are geotagged and organized in a graph.

3.2. Category-level object and scene recognition

The objective in this core part of our research is to learn and recognize quickly and accurately thousands of visual categories, including materials, objects, scenes, and broad classes of temporal events, such as patterns of human activities in picnics, conversations, etc. The current paradigm in the vision community is to model/learn one object category (read 2D aspect) at a time. If we are to achieve our goal, we have to break away from this paradigm, and develop models that account for the tremendous variability in object and scene appearance due to texture, material, viewpoint, and illumination changes within each object category, as well as the complex and evolving relationships between scene elements during the course of normal human activities. Our current work, outlined in detail in section 6.2), focuses on the two problems described next.

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). We have also developed fast and scalable optimization methods for learning the sparse image representations, and developed a software called SPAMS (SPArse Modelling Software) presented in Section 5.1. The work of J. Mairal is summarized in his thesis (Mairal, 2010). The most recent work has focused on developing a general formulation for supervised dictionary learning and investigating methods to learn better mid-level features for recognition.

3.2.2. Category-level object/scene recognition and segmentation

Another significant 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: (i) a graph matching kernel for object categorization, (ii) strongly supervised deformable part-based model for object detection/localization, (iii) a spatial pyramid representation incorporating photographic styles for category-level image classification, (iv) a MRF model for segmentation of text in natural scenes, and (v) algorithms for clustering using convex penalties, and fast approximate energy minimization using graph-cuts.

3.3. Image restoration, manipulation and enhancement

The goal of this part of our research is to develop models, and methods for image/video restoration, manipulation and enhancement. The ability to “intelligently" manipulate the content of images and video is just as essential as high-level content interpretation in many applications: This ranges from restoring old films or removing unwanted wires and rigs from new ones in post production, to cleaning up a shot of your daughter at her birthday party, which is lovely but noisy and blurry because the lights were out when she blew the candles, or editing out a tourist from your Roman holiday video. Going beyond the modest abilities of current “digital zoom" (bicubic interpolation in general) so you can close in on that birthday cake, “deblock" a football game on TV, or turn your favorite DVD into a blue-ray, is just as important.

In this context, we believe there is a new convergence between computer vision, machine learning, and signal processing. For example: The idea of exploiting self-similarities in image analysis, originally introduced in computer vision for texture synthesis applications (Efros and Leung, 1999), is the basis for non-local means (Buades *et al.*, 2005), one of today's most successful approaches to image restoration. In turn, by combining a powerful sparse coding approach to non-local means (Dabov *et al.*, 2007) with modern machine learning techniques for dictionary learning (Mairal *et al.*, 2010), we have obtained denoising and demosaicking results that are the state of the art on standard benchmarks (Mairal *et al.*, 2009).

Our current work, outlined in detail in section 6.3, has focused on sparse epitome based methods, hierarchical coding and dictionary learning for image de-noising and deblurring. In addition, we have also developed a new geometrical model for removing image blur due to camera shake, together with its efficient approximation and extension to deal with saturated pixels.

3.4. 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; Duchenne, Laptev, Sivic, Bach, Ponce, 2009) means that massive amounts of labelled data for training and recognizing action models will at long last be available. Our research agenda in this scientific domain is described below and our recent results are outlined in detail in section 6.4.

3.4.1. 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 and action categories. We are currently extending this work towards exploiting richer textual descriptions of human actions and using them for learning more powerful contextual models of human actions in video.

3.4.2. 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 based on realistic assumptions. We explore the ways of enriching standard bag-of-feature representations with the higher-level information on objects and scenes pre-learned on related tasks. We also aim to capture higher level structural relations between humans, objects and scenes. Along these strands we are particularly investigating long-term temporal relations in the video which, for example, enable reasoning about the depth ordering of objects as well as the temporal ordering actions in dynamical scenes.

3.4.3. Crowd characterization in video

Human crowds are characterized by distinct visual appearance and require appropriate tools for their analysis. In our work we develop generic methods for crowd analysis in video aiming to address multiple tasks such as (i) crowd density estimation and localization, (ii) characterization and recognition of crowd behaviours (e.g a person running against the crowd flow) as well as (iii) detection and tracking of individual people in the crowd. We address the challenge of analyzing crowds under the large variation in crowd density, video resolution and scene structure.

3.4.4. Action recognition in still images

Recognition of human actions is usually addressed in the scope of video interpretation. Meanwhile, common human actions such as "reading a book", "playing a guitar" or "writing notes" also provide a natural description for many still images. Motivated by the potential impact of recognizing actions in still images, we address recognition of human actions in consumer photographs. We have so far studied performance of several state-of-the-art visual recognition methods applied to existing datasets and our newly collected dataset with 968 Flickr images and seven classes of human actions. We have also developed a model of person-object interactions and demonstrated its improved performance for recognition of human actions in still images.

3.4.5. Modeling and recognizing person-object and person-scene interactions.

We have currently started to explore this novel research direction. As mentioned above, we have developed a model of person-object interactions in still images. In addition, we have also investigated the use of human pose as a cue for single-view 3D scene understanding. Our method builds upon recent advances in still-image action recognition and pose estimation, to extract functional and geometric constraints about the scene from people detections. These constraints are then used to improve state-of-the-art single-view 3D scene understanding methods.

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. 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. SPArse Modeling Software (SPAMS)

SPAMS v2.1 was released as open-source software in June 2011 (v1.0 was released in September 2009 and v2.0 in November 2010). It is an optimization toolbox implementing algorithms to address various machine learning and signal processing problems involving

- Dictionary learning and matrix factorization (NMF, sparse PCA, ...)
- Solving sparse decomposition problems with LARS, coordinate descent, OMP, SOMP, proximal methods
- Solving structured sparse decomposition problems (ℓ_1/ℓ_2, ℓ_1/ℓ_∞, sparse group lasso, tree-structured regularization, structured sparsity with overlapping groups,...).

The software and its documentation are available at http://www.di.ens.fr/willow/SPAMS/.

5.2. Non-uniform Deblurring for Shaken and Partially Saturated Images

This is a package of Matlab code for non-blind removal of non-uniform camera shake blur from a single blurry image. The package explicitly deals with images containing some saturated pixels. The algorithm is described in [19]. The package is publicly available at http://www.di.ens.fr/willow/research/saturation/.

5.3. Local dense and sparse space-time features

This is a package with Linux binaries implementing extraction of local space-time features in video. The package was updated in January 2011. The code supports feature extraction at Harris3D points, on a dense space-time grid as well as at user-supplied space-time locations. The package is publicly available at http://www.di.ens.fr/~laptev/download/stip-2.0-linux.zip.

5.4. Segmenting Scenes by Matching Image Composites

This is a package of Matlab code implementing unsupervised data-driven scene segmentation as described in (Russell *et al.* NIPS 2009). The package was created in June 2011 and is available at http://www.cs.washington.edu/homes/bcr/projects/SceneComposites/index.html.

5.5. Discriminative Clustering for Image Co-segmentation

This is a package of Matlab code implementing unsupervised discriminative clustering for co-segmenting multiple images described in (Joulin *et al.* CVPR 2010) and (Joulin *et al.* NIPS 2010). The aim is to segment a given set of images containing objects from the same category, simultanously and without prior information. The package was last updated in October 2011 and is available at http://www.di.ens.fr/~joulin/code/coseg.zip.

5.6. Clustering with Convex Fusion Penalties

This is a package of Matlab code implementing a hierarchical clustering with convex fusion penalties described in (Hocking *et al.* ICML 2011 [10]). The package is available at http://www.di.ens.fr/~joulin/code/clusterpath_norm_Inf.zip.

6. New Results

6.1. 3D object and scene modeling, analysis, and retrieval

6.1.1. Quantitative image analysis for archeology

Participants: Bryan Russell, Jean Ponce, Josef Sivic, Helene Dessales [ENS Archeology laboratory].

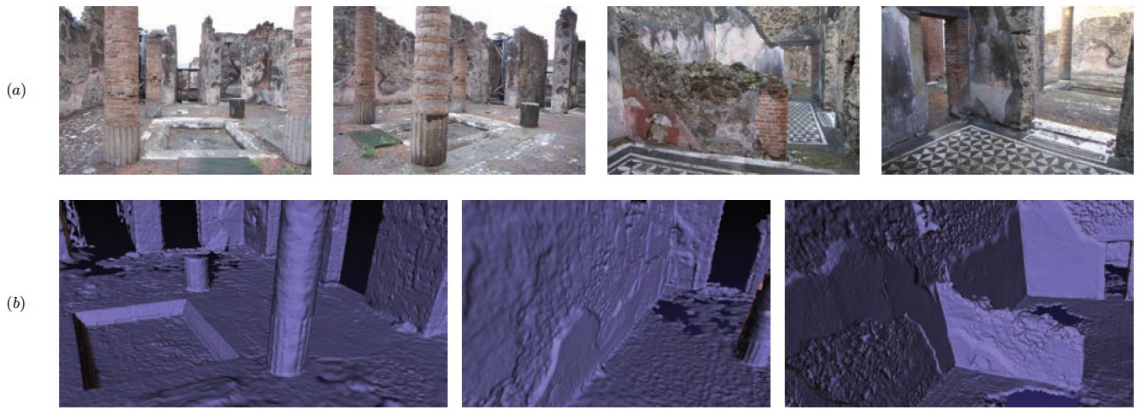

Figure 1. (a) Example photographs captured of the Pompeii site (563 photographs are used in total). (b) Rendered viewpoints of the recovered 3D model. Notice the fine-level details that are captured by the model.

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.

Recently, we have addressed the problem of automatically aligning historical architectural paintings with 3D models obtained using multi-view stereo technology from modern photographs. This is a challenging task because of the variations in appearance, geometry, color and texture due to environmental changes over time, the nonphotorealistic nature of architectural paintings, and differences in the viewpoints used by the painters and photographers. Our alignment procedure consists of two novel aspects: (i) we combine the gist descriptor with the view-synthesis/retrieval of Irschara et al. to obtain a coarse alignment of the painting to the 3D model, and (ii) we have developed an ICP-like viewpoint refinement procedure, where 3D surface orientation discontinuities (folds and creases) and view-dependent occlusion boundaries are rendered from the automatically obtained and noisy 3D model in a view-dependent manner and matched to gPB contours extracted from the paintings. We demonstrate the alignment of XIXth Century architectural watercolors of the Casa di Championnet in Pompeii with a 3D model constructed from modern photographs using the PMVS public-domain multi-view stereo software. Figure 1 shows some of the captured photographs and snapshots of the 3D reconstruction of the site. Notice that the 3D reconstruction captures much detail of the walls and structures. Example painting to 3D model alignments are shown in figure 2.

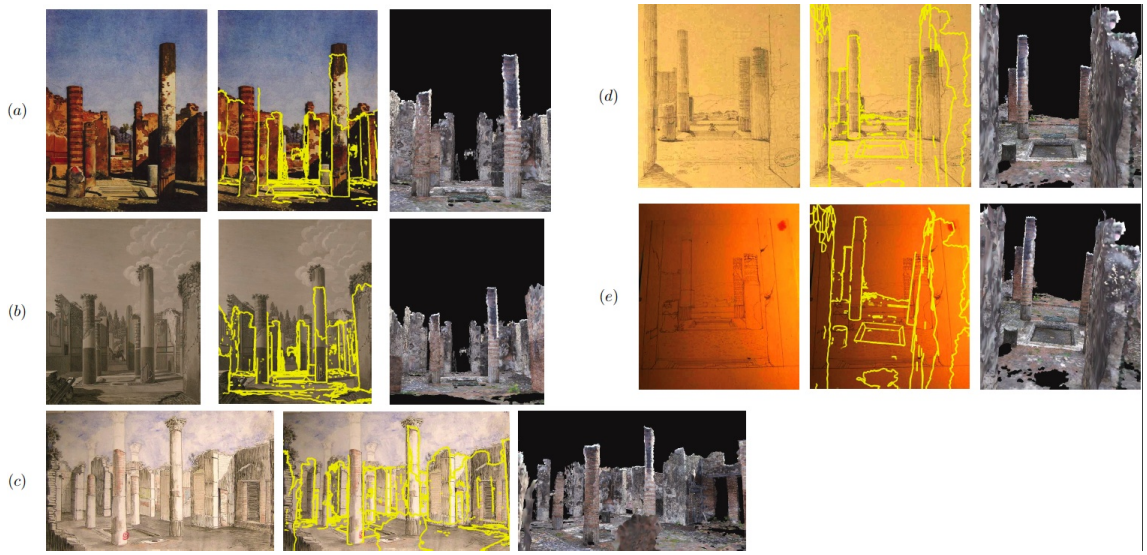

Figure 2. Final alignment between the paintings and 3D model. For each example, left: painting; middle: 3D model contours projected onto painting; right: synthesized viewpoint from 3D model using recovered camera parameters. For the examples in (a-c), note how the final alignment is close to the painting. Our system handles paintings that depict the 3D structure of the scene over time and span different artistic styles and mediums (e.g. water colors, cross-hatching, copies of originals on engravings). Notice how the site changes over time, with significant structural changes (e.g. the wall murals decay over time, the columns change). Example failure cases are shown in (d,e).

This work resulted in a workshop publication [16].

6.1.2. Visual localization by linear combination of image descriptors

Participants: Josef Sivic, Akihiko Torii [Tokyo Institute of Technology], Tomas Pajdla [CTU in Prague].

In this work, we seek to predict the GPS location of a query image given a database of images localized on a map with known GPS locations. The contributions of this work are three-fold: (1) we formulate the image-based localization problem as a regression on an image graph with images as nodes and edges connecting close-by images; (2) we design a novel image matching procedure, which computes similarity between the query and pairs of database images using edges of the graph and considering linear combinations of their fea- ture vec- tors. This improves generalization to unseen viewpoints and illumination conditions, while reducing the database size; (3) we demonstrate that the query location can be pre- dicted by interpolating locations of matched images in the graph without the costly estimation of multi-view geometry. We demonstrate benefits of the proposed image matching scheme on the standard Oxford building benchmark, and show localization results on a database of 8,999 panoramic Google Street View images of Pittsburgh.

This work resulted in a publication [18].

6.2. Category-level object and scene recognition

6.2.1. Task-Driven Dictionary Learning

Participants: Julien Mairal, Jean Ponce, Francis Bach [INRIA SIERRA].

Modeling data with linear combinations of a few elements from a learned dictionary has been the focus of much recent research in machine learning, neuroscience and signal processing. For signals such as natural images that admit such sparse representations, it is now well established that these models are well suited to restoration tasks. In this context, learning the dictionary amounts to solving a large-scale matrix factorization problem, which can be done efficiently with classical optimization tools. The same approach has also been used for learning features from data for other purposes, e.g., image classification, but tuning the dictionary in a supervised way for these tasks has proven to be more difficult. In this paper, we present a general formulation for supervised dictionary learning adapted to a wide variety of tasks, and present an efficient algorithm for solving the corresponding optimization problem. Experiments on handwritten digit classification, digital art identification, nonlinear inverse image problems, and compressed sensing demonstrate that our approach is effective in large-scale settings, and is well suited to supervised and semi-supervised classification, as well as regression tasks for data that admit sparse representations.

This work has resulted in a publication [4].

6.2.2. Ask the locals: multi-way local pooling for image recognition

Participants: Y-Lan Boureau, Jean Ponce, Nicolas Le Roux [INRIA SIERRA], Francis Bach [INRIA SIERRA], Yann LeCun [New York University].

Invariant representations in object recognition systems are generally obtained by pooling feature vectors over spatially local neighborhoods. But pooling is not local in the feature vector space, so that widely dissimilar features may be pooled together if they are in nearby locations. Recent approaches rely on sophisticated encoding methods and more specialized codebooks (or dictionaries), e.g., learned on subsets of descriptors which are close in feature space, to circumvent this problem. In this work, we argue that a common trait found in much recent work in image recognition or retrieval is that it leverages locality in feature space on top of purely spatial locality. We propose to apply this idea in its simplest form to an object recognition system based on the spatial pyramid framework, to increase the performance of small dictionaries with very little added engineering. Stateof- the-art results on several object recognition benchmarks show the promise of this approach.

This work has resulted in a publication [7].

6.2.3. A Graph-matching Kernel for Object Categorization

Participants: Olivier Duchenne, Armand Joulin, Jean Ponce.

This paper addresses the problem of category-level im- age classification. The underlying image model is a graph whose nodes correspond to a dense set of regions, and edges reflect the underlying grid structure of the image and act as springs to guarantee the geometric consistency of nearby regions during matching. A fast approximate algorithm for matching the graphs associated with two images is pre- sented. This algorithm is used to construct a kernel appro- priate for SVM-based image classification, and experiments with the Caltech 101, Caltech 256, and Scenes datasets demonstrate performance that matches or exceeds the state of the art for methods using a single type of features.

This work has resulted in an ICCV 2011 publication [9] (oral presentation).

6.2.4. A Tensor-Based Algorithm for High-Order Graph Matching

Participants: Olivier Duchenne, Jean Ponce, Francis Bach [INRIA SIERRA], Inso Kweon [KAIST, Korea].

This paper addresses the problem of establishing correspondences between two sets of visual features using higher-order constraints instead of the unary or pairwise ones used in classical methods. Concretely, the corresponding hypergraph matching problem is formulated as the maximization of a multilinear objective function over all permutations of the features. This function is defined by a tensor representing the affinity between feature tuples. It is maximized using a generalization of spectral techniques where a relaxed problem is first solved by a multi-dimensional power method, and the solution is then projected onto the closest assignment matrix. The proposed approach has been implemented, and it is compared to state-of-the-art algorithms on both synthetic and real data.

This work has resulted in an PAMI publication [2].

6.2.5. Clusterpath: an algorithm for clustering using convex fusion penalties

Participants: Armand Joulin, Toby Hocking [INRIA SIERRA], Francis Bach [INRIA SIERRA], Jean-Philippe Vert [Mines ParisTech].

We present a new clustering algorithm by proposing a convex relaxation of hierarchical clustering, which results in a family of objective functions with a natural geometric interpretation. We give efficient algorithms for calculating the continuous regularization path of solutions, and discuss relative advantages of the parameters. Our method experimentally gives state-ofthe-art results similar to spectral clustering for non-convex clusters, and has the added benefit of learning a tree structure from the data.

This work has resulted in an publication [10].

6.2.6. An MRF model for binarization of natural scene text

Participants: Karteek Alahari, Anand Mishra [IIT India], C.V. Jawahar [IIT India].

Scene text recognition has gained significant attention from the computer vision community in recent years. Recognizing text in the wild is a challenging problem, even more so than the recognition of scanned documents. In this work, we focus on the problem of cropped word recognition. We present a framework that exploits both bottom-up and top-down cues. The bottom-up cues are derived from individual character detections from the image. We build a Conditional Random Field model on these detections to jointly model the strength of the detections and the interactions between them. We impose top-down cues obtained from a lexicon-based prior, i.e. language statistics, on the model. The optimal word represented by the text image is obtained by minimizing the energy function corresponding to the random field model.

We show very significant improvements in accuracies on two challenging public datasets, namely Street View Text (over 15%) and ICDAR 2003 (over 10%).

This work has resulted in an publication [12].

6.2.7. Strongly-supervised deformable part model for object detection

Participants: Hossein Azizpour [KTH Stockholm], Ivan Laptev, Stefan Carlsson [KTH Stockholm].

Deformable part models achieve state-of-the-art performance for object detection while relying on the greedy initialization during training. The goal of this paper is to investigate limitations of such initialization and to improve the model for the case when part locations are known at the training time. To this end, we deploy part-level supervision and demonstrate improved detection results when learning models with manually-initialized part locations. We further explore the benefits of the strong supervision and learn model structure by minimizing the variance among adjacent model parts. Our method can simultaneously handle samples with and without part-level annotation making benefit even from a fraction of fully-annotated training samples. Experimental results are reported for the detection of six animal classes in PASCAL VOC 2007 and 2010 datasets. We demonstrate significantly improved performance of our model compared to the state-of-the-art LSVM object detector and poselet detector. Example learnt models are shown in figure 3.

This work has resulted in a submission to CVPR 2012.

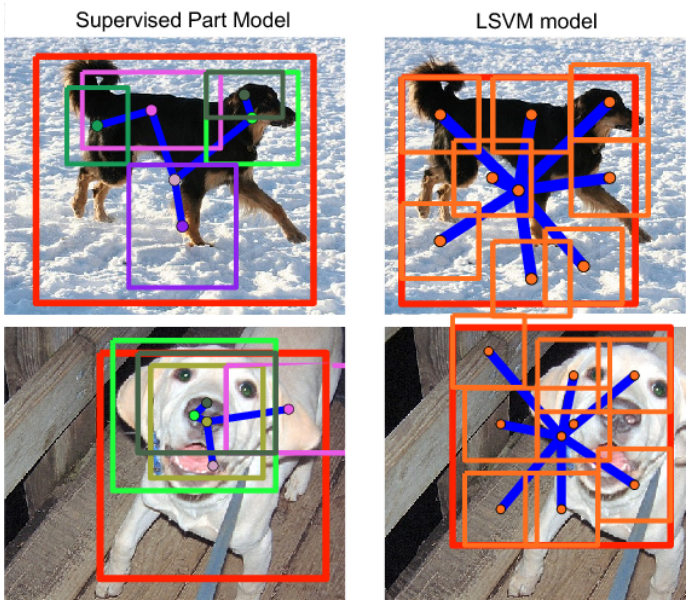

Figure 3. Illustration of deformable part models trained for the dog class. Left: dog detection with the proposed supervised part model. Right: dog detection with the original LSVM model of Felzenszwalb et al. Our model adapts to the different object appearance while LSVM model attempts to explain both samples using the same deformable HOG template.

6.2.8. Exploiting Photographic Style for Category-Level Image Classification by Generalizing the Spatial Pyramid

Participant: Jan van Gemert [University of Amsterdam].

This paper investigates the use of photographic style for category-level image classification. Specifi

cally, we exploit the assumption that images within a category share a similar style defi

ned by attributes such as colorfulness, lighting, depth of fi

eld, viewpoint and saliency. For these style attributes we create correspondences across images by a generalized spatial pyramid matching scheme. Where the spatial pyramid groups features spatially, we allow more general feature grouping and in this paper we focus on grouping images on photographic style. We evaluate

our approach in an object classification task and investigate style differences between professional and amateur photographs. We show that a generalized pyramid with style-based attributes improves performance on the professional Corel and amateur Pascal VOC 2009 image datasets.

This work has resulted in a publication [20].

6.2.9. *Generalized Fast Approximate Energy Minimization via Graph Cuts: Alpha-Expansion Beta-Shrink Moves*

Participants: Karteek Alahari, Mark Schmidt [INRIA SIERRA].

We present alpha-expansion beta-shrink moves, a simple generalization of the widely-used alpha beta-swap and alpha-expansion algorithms for approximate energy minimization. We show that in a certain sense, these moves dominate both alpha beta-swap and alpha-expansion moves, but unlike previous generalizations the new moves require no additional assumptions and are still solvable in polynomial-time. We show promising experimental results with the new moves, which we believe could be used in any context where alpha-expansions are currently employed.

This work has resulted in a publication [17].

6.3. Image restoration, manipulation and enhancement

6.3.1. *Non-uniform Deblurring for Shaken Images*

Participants: Oliver Whyte, Josef Sivic, Andrew Zisserman, Jean 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 significantly 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 first 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.

The project resulted in a publication [5].

6.3.2. *Deblurring shaken and partially saturated images*

Participants: Oliver Whyte, Josef Sivic, Andrew Zisserman.

We address the problem of deblurring images degraded by camera shake blur and saturated or over-exposed pixels. Saturated pixels are a problem for existing non-blind deblurring algorithms because they violate the assumption that the image formation process is linear, and often cause significant artifacts in deblurred outputs. We propose a forward model that includes sensor saturation, and use it to derive a deblurring algorithm properly treating saturated pixels. By using this forward model and reasoning about the causes of artifacts in the deblurred results, we obtain significantly better results than existing deblurring algorithms. Further we propose an efficient approximation of the forward model leading to a significant speed-up. Example result is shown in figure 4.

The project resulted in a publication [19].

6.3.3. *Dictionary Learning for Deblurring and Digital Zoom*

Participants: Florent Couzinie, Julien Mairal, Jean Ponce, Francis Bach [INRIA SIERRA].

(a) Blurry image (b) Deblurred with Richardson-Lucy (c) Deblurred with algorithm of Krishnan & Fergus [11] (d) Deblurred with proposed method

Figure 4. Deblurring saturated images. Note that the ringing around saturated regions, visible in columns (b) and (c) is removed by our method (d), without causing any loss in visual quality elsewhere.

This work proposes a novel approach to image deblurring and digital zooming using sparse local models of image appearance. These models, where small image patches are represented as linear combinations of a few elements drawn from some large set (dictionary) of candidates, have proven well adapted to several image restoration tasks. A key to their success has been to learn dictionaries adapted to the reconstruction of small image patches. In contrast, recent works have proposed instead to learn dictionaries which are not only adapted to data reconstruction, but also tuned for a specific task. We introduce here such an approach to deblurring and digital zoom, using pairs of blurry/sharp (or low-/high-resolution) images for training, as well as an e?ective stochastic gradient algorithm for solving the corresponding optimization task. Although this learning problem is not convex, once the dictionaries have been learned, the sharp/high-resolution image can be recovered via convex optimization at test time. Experiments with synthetic and real data demonstrate the e?ectiveness of the proposed approach, leading to stateof-the-art performance for non-blind image deblurring and digital zoom.

This work has resulted in a publication [1].

6.3.4. Sparse Image Representation with Epitomes

Participants: Louise Benoit, Julien Mairal, Jean Ponce, Francis Bach [INRIA SIERRA].

Sparse coding, which is the decomposition of a vector using only a few basis elements, is widely used in machine learning and image processing. The basis set, also called dictionary, is learned to adapt to specific data. This approach has proven to be very effective in many image processing tasks. Traditionally, the dictionary is an unstructured "flat" set of atoms. In this work, we study structured dictionaries which are obtained from an epitome, or a set of epitomes. The epitome is itself a small image, and the atoms are all the patches of a chosen size inside this image. This considerably reduces the number of parameters to learn and provides sparse image decompositions with shift invariance properties. We propose a new formulation and an algorithm for learning the structured dictionaries associated with epitomes, and illustrate their use in image denoising tasks.

This work has resulted in a CVPR'11 publication [6].

6.3.5. Proximal Methods for Hierarchical Sparse Coding

Participants: Julien Mairal, Rodolphe Jenatton [INRIA SIERRA], Guillaume Obozinski [INRIA SIERRA], Francis Bach [INRIA SIERRA].

Sparse coding consists in representing signals as sparse linear combinations of atoms selected from a dictionary. We consider an extension of this framework where the atoms are further assumed to be embedded in a tree. This is achieved using a recently introduced tree-structured sparse regularization norm, which has proven useful in several applications. This norm leads to regularized problems that are difficult to optimize, and in this paper, we propose efficient algorithms for solving them. More precisely, we show that the proximal operator associated with this norm is computable exactly via a dual approach that can be viewed as the composition of elementary proximal operators. Our procedure has a complexity linear, or close to linear, in the number of atoms, and allows the use of accelerated gradient techniques to solve the tree-structured sparse approximation problem at the same computational cost as traditional ones using the l1-norm. Our method is efficient and scales gracefully to millions of variables, which we illustrate in two types of applications: first, we consider fixed hierarchical dictionaries of wavelets to denoise natural images. Then, we apply our optimization tools in the context of dictionary learning, where learned dictionary elements naturally self-organize in a prespecified arborescent structure, leading to better performance in reconstruction of natural image patches. When applied to text documents, our method learns hierarchies of topics, thus providing a competitive alternative to probabilistic topic models.

This work has resulted in a publication [3].

6.4. Human activity capture and classification

6.4.1. Track to the future: Spatio-temporal video segmentation with long-range motion cues

Participants: Jose Lezama, Karteek Alahari, Ivan Laptev, Josef Sivic.

Video provides rich visual cues such as motion and appearance but also much less explored long-range temporal interactions among objects. We aim to capture such interactions and to construct powerful intermediate-level video representation for subsequent recognition. Motivated by this goal, we seek to obtain spatio-temporal oversegmentation of the video into regions that respect object boundaries and, at the same time, associate object pixels over many video frames. The contributions of this paper are twofold. First, we develop an efficient spatio-temporal video segmentation algorithm, that naturally incorporates long-range motion cues from the past and future frames in the form of clusters of point tracks with coherent motion. Second, we devise a new track clustering cost-function that includes occlusion reasoning, in the form of depth ordering constraints, as well as motion similarity along the tracks. We evaluate the proposed approach on a challenging set of video sequences of office scenes from feature length movies.

This work resulted in a publication [11].

6.4.2. Density-aware person detection and tracking in crowds

Participants: Mikel Rodriguez, Ivan Laptev, Josef Sivic, Jean-Yves Audibert [INRIA SIERRA].

We address the problem of person detection and tracking in crowded video scenes. While the detection of individual objects has been improved significantly over the recent years, crowd scenes remain particularly challenging for the detection and tracking tasks due to heavy occlusions, high person densities and significant variation in people's appearance. To address these challenges, we propose to lever- age information on the global structure of the scene and to resolve all detections jointly. In particular, we explore con- straints imposed by the crowd density and formulate per- son detection as the optimization of a joint energy function combining crowd density estimation and the localization of individual people. We demonstrate how the optimization of such an energy function significantly improves person de- tection and tracking in crowds. We validate our approach on a challenging video dataset of crowded scenes. The proposed approach is illustrated in figure 5.

This work has resulted in a publication [14].

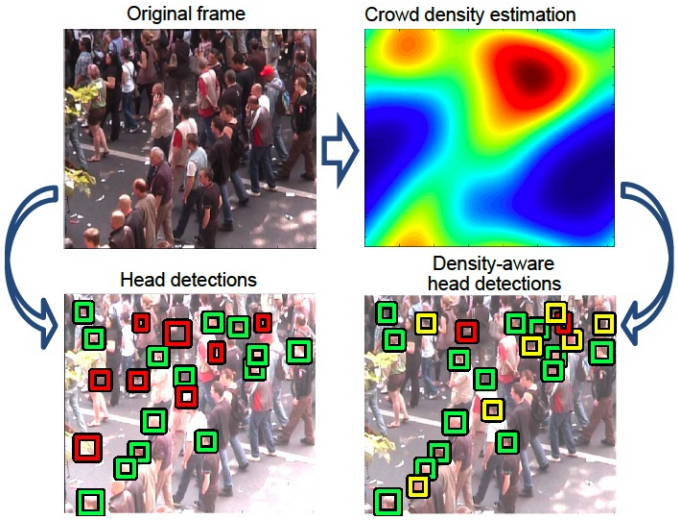

Figure 5. Individual head detections provided by state-of-the-art object detector (Felzenswalb et al. 2009) (bottom-left; green: true positives; red: false positives) are improved significantly by our method (bottom-right; yellow: new true positives) using the crowd density estimate (topright) obtained from the original frame (top-left).

6.4.3. Data-driven Crowd Analysis in Videos

Participants: Mikel Rodriguez, Josef Sivic, Ivan Laptev, Jean-Yves Audibert [INRIA SIERRA].

In this work we present a new crowd analysis algorithm powered by behavior priors that are learned on a large database of crowd videos gathered from the Internet. The algorithm works by first learning a set of crowd behavior priors off-line. During testing, crowd patches are matched to the database and behavior priors are transferred. We ad- here to the insight that despite the fact that the entire space of possible crowd behaviors is infinite, the space of distin- guishable crowd motion patterns may not be all that large. For many individuals in a crowd, we are able to find anal- ogous crowd patches in our database which contain sim- ilar patterns of behavior that can effectively act as priors to constrain the difficult task of tracking an individual in a crowd. Our algorithm is data-driven and, unlike some crowd characterization methods, does not require us to have seen the test video beforehand. It performs like state-of- the-art methods for tracking people having common crowd behaviors and outperforms the methods when the tracked individual behaves in an unusual way.

This work has resulted in a publication [15].

6.4.4. Learning person-object interactions for action recognition in still images

Participants: Vincent Delaitre, Josef Sivic, Ivan Laptev.

In this work, we investigate a discriminatively trained model of person-object interactions for recognizing common human actions in still images. We build on the locally order-less spatial pyramid bag-of-features model, which was shown to perform extremely well on a range of object, scene and human action recognition tasks. We introduce three principal contributions. First, we replace the standard quan- tized local HOG/SIFT features with stronger discriminatively trained body part and object detectors. Second, we introduce new person-object interaction features based on spatial co-occurrences of individual body parts and objects. Third, we address the combinatorial problem of a large number of possible interaction pairs and propose a discriminative selection procedure using a linear support vector machine (SVM) with a sparsity inducing regularizer. Learning of action-specific body part and object interactions bypasses the difficult problem of estimating the complete human body pose configuration. Benefits of the proposed model are shown on human

action recognition in consumer photographs, outperforming the strong bag-of-features baseline. The proposed model is illustrated in figure 6.

This work has resulted in a publication [8].

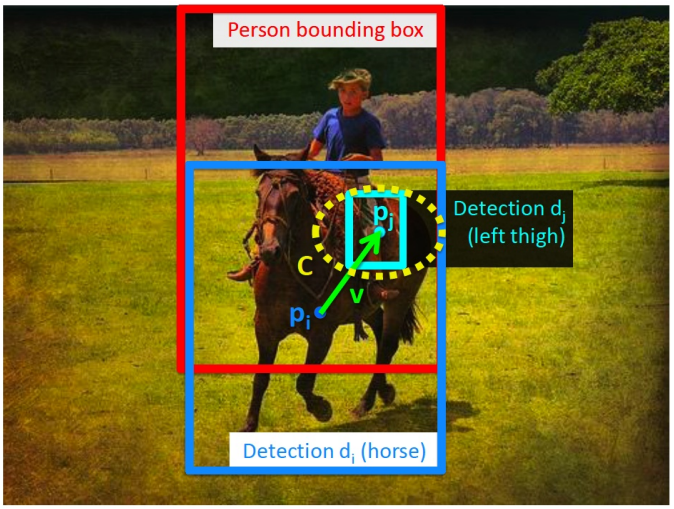

Figure 6. Representing person-object interactions by pairs of body part (cyan) and object (blue) detectors. To get a strong interaction response, the pair of detectors (here visualized at positions pi and pj) must fire in a particular relative 3D scale-space displacement (given by the vector v) with a scale-space displacement uncertainty (deformation cost) given by diagonal 3x3 covariance matrix C (the spatial part of C is visualized as a yellow dotted ellipse). Our image representation is defined by the max-pooling of interaction responses over the whole image, solved efficiently by the distance transform.

6.4.5. People Watching: Human Actions as a Cue for Single View Geometry

Participants: David Fouhey [CMU], Vincent Delaitre, Abhinav Gupta [CMU], Ivan Laptev, Alexei Efros [CMU], Josef Sivic.

We present an approach which exploits the coupling between human actions and scene geometry. We investigate the use of human pose as a cue for single-view 3D scene understanding. Our method builds upon recent advances in still-image action recognition and pose estimation, to extract functional and geometric constraints about the scene from people detections. These constraints are then used to improve state-of-the-art single-view 3D scene understanding approaches. The proposed method is validated on a collection of single-viewpoint time-lapse image sequences as well as a dataset of still images of indoor scenes. We demonstrate that observing people performing different actions can significantly improve estimates of scene geometry and 3D layout. The main idea of this work is illustrated in figure 7.

This work is in submission to CVPR 2012.

6.4.6. Joint pose estimation and action recognition in image graphs

Participants: K. Raja [INRIA Rennes], Ivan Laptev, Patrick Perez [Technicolor], L. Osei [INRIA Rennes].

Human analysis in images and video is a hard problem due to the large variation in human pose, clothing, camera view-points, lighting and other factors. While the explicit modeling of this variability is difficult, the huge amount of available person images motivates for the implicit, datadriven approach to human analysis. In this work we aim to explore this approach using the large amount of images spanning a subspace of human appearance. We model this subspace by connecting images into a graph and propagating information through

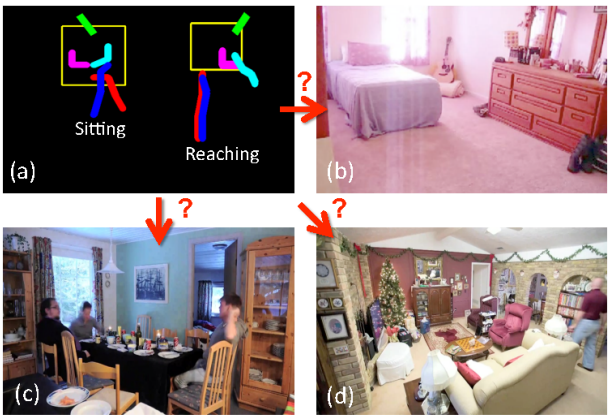

Figure 7. What can human actions tell us about the 3D structure of the scene? Quite a lot, actually. Consider the two person detections and their estimated pose in (a). They were detected in a time-lapse sequence of one of the three scenes (b-d). Can you guess which one? Most people can easily see that it is (b). Even though this is only a static image, the actions and the pose of the disembodied figures reveal a lot about the geometric structure of the scene. The pose of the left fig- ure reveals a horizontal surface right under its pelvis, which ends abruptly at the knees. The right figure's pose reveals a ground plane under its feet as well as a likely horizontal surface near the hand location. In both cases we observe a strong physical and functional coupling that exists between people and the 3D geometry of the scene. Our aim in this work is to exploit this coupling.

such a graph using a discriminatively trained graphical model. We particularly address the problems of human pose estimation and action recognition and demonstrate how image graphs help solving these problems jointly. We report results on still images with human actions from the KTH dataset.

This work has resulted in a publication [13].

6.5. Creation of the SIERRA project-team

6.5.1. From WILLOW alone to WILLOW and SIERRA

The WILLOW team officially started in the Spring of 2007. From the start, it was clear that machine learning was a key ingredient to new breakthroughs, and our activities have steadily grown in this area. In three short years, WILLOW has grown into a mature group of about 30 people, and it divides its activities between computer vision, machine learning, and the cross-pollination of the two fields, with video as one of the core research areas. We have been very successful, with many publications in all the major international conferences and leading journals in both areas, but we are a large group with very diverse interests, ranging from camera geometry to statistics, and from image retrieval to bioinformatics applications of structured sparse coding. With the creation of the SIERRA project-team, the core machine learning activities of WILLOW have been be transferred to the new group.

The two teams continue collaborating with each other (they remain co-located at the INRIA site in central Paris), but have a sharper focus on their respective computer vision and machine learning activities.

6.5.2. SIERRA

The SIERRA project-team was created by the INRIA on January 1st 2011 and is headed by Francis Bach, who received in 2009 a Jr. ERC grant.

7. Contracts and Grants with Industry

7.1. EADS (ENS)

Participants: Jean Ponce, Josef Sivic, Andrew Zisserman.

The WILLOW team has had collaboration efforts with EADS via tutorial presentations and discussions with A. Zisserman, J. Sivic and J. Ponce at EADS and ENS, and submitting joint grant proposals. In addition, Marc Sturzel (EADS) is doing a PhD at ENS with Jean Ponce and Andrew Zisserman.

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

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

This collaborative project, already mentioned several times in this report, brings together the WILLOW and LEAR 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.

7.3. DGA: CrowdChecker (ENS and E-vitech)

Participants: Jean Ponce, Josef Sivic, Ivan Laptev.

CrowdChecker (DGA) is a joint DGA project with industrial partner E-vitech. This contract belongs to our video understanding research program. It aims at real-time characterization of a crowd seen from a camera mounted 3 to 10 meters over the ground. It includes segmentation of the crowd, clustering by movement, detection of abnormal behaviors (persons, for instance, crossing the crowd flow, or having unusual speed), tracking people. Several parts of computer vision and machine learning are involved: crowd optical flow estimation, image processing, crowd feature extraction, statistical learning from video database, etc.

7.4. PersonSpace (INRIA and Technicolor-R&D)

Participant: Ivan Laptev.

PersonSpace is a CIFRE PhD contract with Technicolor-R&D. The project addresses the problem of human pose estimation and human action recognition in still images. We investigate a subspace spanned by images and videos of people and explore the structure of this subspace to formulate useful constraints for automatic interpretation of person images.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. Agence Nationale de la Recherche: DETECT (ENS)

Participant: 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 and F. Bach from INRIA SIERRA team and J. Sivic from Willow.

S. Arlot (INRIA SIERRA) is the leader of this ANR "Young researchers" project.

8.2. European Initiatives

8.2.1. QUAERO (INRIA)

Participant: Ivan Laptev.

QUAERO (AII) is a European collaborative research and development program with the goal of developing multimedia and multi-lingual indexing and management tools for professional and public applications. Quaero consortium involves 24 academic and industrial partners leaded by Technicolor (previously Thomson). Willow participates in work package 9 "Video Processing" and leads work on motion recognition and event recognition tasks.

8.2.2. EIT-ICT: Cross-linking Visual Information and Internet Resources using Mobile Networks (INRIA)

Participants: Ivan Laptev, Josef Sivic.

The goal of this project within the European EIT-ICT activity is to perform basic research in the area of semantic image and video understanding as well as efficient and reliable indexing into visual databases with a specific focus on indexing visual information captured by mobile users into Internet resources. The aim is demonstrate future applications and push innovation in the field of mobile visual search.

This is a collaborative effort with C. Schmid (INRIA Grenoble) and S. Carlsson (KTH Stockholm).

8.2.3. European Research Council (ERC) Advanced Grant

Participants: Jean Ponce, Ivan Laptev, Josef Sivic.

WILLOW will be funded in part from 2011 to 2015 by the ERC Advanced Grant "VideoWorld" awarded to Jean Ponce by the European Research Council.
This project is concerned with the automated computer analysis of video streams: Digital video is everywhere, at home, at work, and on the Internet. Yet, effective technology for organizing, retrieving, improving, and editing its content is nowhere to be found. Models for video content, interpretation and manipulation inherited from still imagery are obsolete, and new ones must be invented. With a new convergence between computer vision, machine learning, and signal processing, the time is right for such an endeavor. Concretely, we will develop novel spatio-temporal models of video content learned from training data and capturing both the local appearance and nonrigid motion of the elements—persons and their surroundings—that make up a dynamic scene. We will also develop formal models of the video interpretation process that leave behind the architectures inherited from the world of still images to capture the complex interactions between these elements, yet can be learned effectively despite the sparse annotations typical of video understanding scenarios. Finally, we will propose a unified model for video restoration and editing that builds on recent advances in sparse coding and dictionary learning, and will allow for unprecedented control of the video stream. This project addresses fundamental research issues, but its results are expected to serve as a basis for groundbreaking technological advances for applications as varied as film post-production, video archival, and smart camera phones.

9. Dissemination

9.1. Animation of the scientific community

+ Conference and workshop organization
  - I. Laptev, Co-organizer of the Workshop on Gesture Recognition at CVPR 2011. http://clopinet.com/isabelle/Projects/CVPR2011
  - A. Zisserman, Co-organizer of the PASCAL Visual Object Classes Challenge 2011 (VOC2011). http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2011/
  - A. Zisserman, Co-organizer of the PASCAL VOC 2011 workshop at ICCV 2011 http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2011/workshop/index.html
  - A. Zisserman, Co-organizer of the Mysore Park Workshop on Computer Vision 2011
+ Editorial Boards
  - International Journal of Computer Vision (I. Laptev, J. Ponce, J. Sivic and A. Zisserman).
  - Image and Vision Computing Journal (I. Laptev).
  - Foundations and Trends in Computer Graphics and Vision (J. Ponce, A. Zisserman).
  - SIAM Journal on Imaging Sciences (J. Ponce)
+ Area Chairs
  - IEEE Conference on Computer Vision and Pattern Recognition, 2011 (J. Sivic).
  - IEEE International Conference on Automatic Face and Gesture Recognition, 2011 (I. Laptev).
  - IEEE International Conference on Computer Vision, 2011 (I. Laptev, J. Sivic, A. Zisserman (Program board member)).
  - European Conference on Computer Vision, 2012 (I. Laptev, J. Ponce, J. Sivic, A. Zisserman).

+ Program Committees
  - IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011 (I. Laptev, A. Zisserman, M. Rodriguez).
  - IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012 (I. Laptev, J. Sivic, A. Zisserman)
  - Reviewer for the ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH ASIA), 2011 (J. Sivic, A. Zisserman).
  - IEEE International Conference on Computer Vision, 2011 (K. Alahari, O. Duchenne, M. Rodriguez, O. Whyte).
  - International Conference on Neural Information Processing Systems (NIPS), 2011 (A. Zisserman).
  - International Conference on Robotics and Automation (ICRA), 2011 (J. Sivic, A. Zisserman).
  - A. Zisserman, reviewer for ERC grant applications.

+ PhD thesis committee:
  - Biliana Kaneva, MIT, 2011 (J. Sivic)
  - Meriem Bendris, Telecom ParisTech, 2011 (J. Sivic)
  - Jerome Revaud, INSA Lyon (J. Ponce)
  - Sebastien Dalibard, LAAS Toulouse (J. Ponce)
  - Hedi Harzallah INRIA Rhone-Alpes (J. Ponce)
  - Amael Delaunoy INRIA Rhone-Alpes (J. Ponce)
  - Hoang Hiep Vu, ENPC (J. Ponce)
  - Ana Lopes, Federal University of Minas Gerais, Brasil, 2011 (I. Laptev)
  - Pyry Matikainen, CMU, USA, 2011 (I. Laptev)
  - Shuji Zhao, ETIS/LIP6, France, 2011 (I. Laptev)
  - Sibt ul Hussain, Laboratoire Jean Kuntzmann, Grenoble (A. Zisserman)

+ HDR thesis committee:
  - Theodore Papadopoulo, Universite de Nice. (J. Ponce)

+ Prizes:
  - Andrew Zisserman was awarded the Rank Prize for his “outstanding contributions to modern computer vision".
  - Julien Mairal, a former PhD student of J. Ponce and F. Bach wins several prizes for his PhD thesis about "Sparse coding for machine learning, image processing and computer vision", http://www.di.ens.fr/~mairal/resources/pdf/phd_thesis.pdf :
    - Best Thesis 2011 Prize of Information and Communication Sciences and Technologies of the EADS foundation. http://ns365501.ovh.net/en/2011-best-thesis-prize-winners
    - runner-up for the 2011 Gilles-Kahn prize (computer science). Prize awarded by Specif (society of computer science researchers and teachers of France) and supervised by the Academie des Sciences. http://www.specif.org/prix-these/historique.html
    - the 2010 PhD thesis prize from AFRIF (pattern recognition) http://www.afrif.asso.fr/node/12
  - INRIA Prime d’excellence scientifique (I. Laptev, J. Sivic)

+ Other:
  - Jean Ponce was named Head of the ENS computer science department in September 2011.
  - Jean Ponce was named Head of the ENS computer science laboratory (LIENS, joint ENS/CNRS/Inria UMR 8548) in September 2011.
  - Jean Ponce was in charge of the LIENS activities linked to the French “investissements d’avenir” (“investing in the future”) initiative, including the LIENS parts of two successful projects, the Laboratory of Excellence (“LABEX”) project of the Foundation of Mathematical Sciences, and the Excellence Initiative (“IDEX”) project of the “Paris Sciences et Lettres” Foundation, granted in 2011.
  - Comites de selection de l’Universite de Caen, May 2011 (I. Laptev)

9.2. Teaching

- J. Ponce, "Introduction to computer vision", L3, Ecole normale supérieure, 36h.
- M. Pocchiola and J. Ponce, "Geometric bases of computer science, L3, Ecole normale supérieure, 36h.
- I. Laptev, J. Ponce and J. Sivic (together with C. Schmid (INRIA Grenoble)), “Object recognition and computer vision”, M2, Ecole normale supérieure, and MVA, Ecole normale supérieure de Cachan, 36h.
- I. Laptev and J. Sivic were speakers at the International Computer Vision Summer School 2011, Sicily, Italy, July 2011.
- I. Laptev and J. Sivic were speakers at the Computer Vision Winter School, ENS Lyon, January 2011.
- A. Zisserman, Optimization lectures (Hilary term 2011), University of Oxford, http://www.robots.ox.ac.uk/~az/lectures/opt/2011/.
- A. Zisserman, Machine learning lectures (Hilary term 2011), University of Oxford, http://www.robots.ox.ac.uk/~az/lectures/ml/2011/.
- A. Zisserman, Teaching on two topics at the INRIA Visual Recognition and Machine Learning Summer School http://www.di.ens.fr/willow/events/cvml2011/.
- I. Laptev and A. Zisserman were speakers at Microsoft Computer Vision School, Moscow, Russia, August 2011 http://summerschool2011.graphicon.ru/en/courses.

9.3. ENS/INRIA Visual Recognition and Machine Learning Summer School 2011

http://www.di.ens.fr/willow/events/cvml2011/

I. Laptev, J. Ponce and J. Sivic (together with C. Schmid (INRIA Grenoble)) co-organized a one week summer school on Visual Recognition and Machine Learning. The summer school, hosted by ENS Ulm, attracted 175 participants from 28 countries (31% France / 48% Europe / 21% other countries (including USA, India, Brazil, Canada, Russia, Japan and China)), and included Master students, PhD students as well as Post-docs and researchers. The summer school provided an overview of the state of the art in visual recognition and machine learning. Lectures were given by 12 speakers (4 USA, 1 UK, 1 Austria, 1 France, 5 INRIA / ENS), which included top international experts in the area of visual recognition (J. Malik, UC Berkeley, USA; M. Hebert and A. Efros, CMU, USA; A. Zisserman, Oxford, UK / WILLOW; L. Bottou, Microsoft, USA). Lectures were complemented by practical sessions to provide participants with hands-on experience with the discussed material. In addition, a poster session was organized for participants to present their current research.

The third summer school in this series is currently in preparation for 2012 to be hosted by INRIA Grenoble.

9.4. Invited presentations

- J. Sivic, Microsoft Research Redmond, USA, Host: R. Szeliski, March 2011
- J. Sivic, Carnegie Mellon University, USA, Host: A. Efros, February 2011
- J. Sivic, AViRS workshop on video surveillance, Paris, Host: D. Marraud, February 2011
- J. Sivic, Czech Technical University in Prague, Host: J. Matas, April 2011
- J. Ponce, Distinguished speaker, Taiwan Academica Sinica, 2011
- J. Ponce, Distinguished speaker, University of Delaware Computer Science Department, 2011
- J. Ponce, ETH Zurich Computer Science Department, 2011
- I. Laptev, Royal Institute of Technology, Stockholm, Sweden, Host: S. Carlsson, December 2011
- I. Laptev, ICCV2011 International Workshop on Video Event Categorisation, Barcelona, Spain, November 2011
- I. Laptev, Assemblee generale du GdR ISIS, Saint-Georges-de-Didonne, France, May 2011
- A. Zisserman, Frontiers workshop on Computer Vision, MIT, August, 2011.
- A. Zisserman, BBC Faces Workshop, September, 2011.

10. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journal

[1] F. COUZINIE, J. MAIRAL, F. BACH, J. PONCE. *Dictionary Learning for Deblurring and Digital Zoom*, in "International Journal of Computer Vision", 2011, In submission, pre-print ArXiv.

[2] O. DUCHENNE, F. BACH, I.-S. KWEON, J. PONCE. *A tensor-based algorithm for high-order graph matching.*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2011, vol. 33, n^o 12, p. 2383–2395.

[3] R. JENATTON, J. MAIRAL, G. OBOZINSKI, F. BACH. *Proximal Methods for Hierarchical Sparse Coding*, in "Journal of Machine Learning Research", July 2011, n^o 12, p. 2297-2334, http://hal.inria.fr/inria-00516723/en.

[4] J. MAIRAL, F. BACH, J. PONCE. *Task-Driven Dictionary Learning*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2011, to appear. preprint ArXiv:1009.5358.

[5] O. WHYTE, J. SIVIC, A. ZISSERMAN, J. PONCE. *Non-uniform Deblurring for Shaken Images*, in "International Journal of Computer Vision", 2011.

International Conferences with Proceedings

[6] L. BENOÎT, J. MAIRAL, F. BACH, J. PONCE. *Sparse Image Representation with Epitomes*, in "Computer Vision and Pattern Recognition", Colorado Springs, United States, June 2011, http://hal.inria.fr/hal-00631652/en.

[7] Y-L. BOUREAU, N. LE ROUX, F. BACH, J. PONCE, Y. LECUN. *Ask the locals: Multi-way local pooling for image recognition*, in "International Conference on Computer Vision", 2011.

[8] V. DELAITRE, J. SIVIC, I. LAPTEV. *Learning person-object interactions for action recognition in still images*, in "Advances in Neural Information Processing Systems", 2011.

[9] O. DUCHENNE, A. JOULIN, J. PONCE. *A Graph-Matching Kernel for Object Categorization*, in "International Conference on Computer Vision", 2011.

[10] T. HOCKING, A. JOULIN, F. BACH, J. VERT. *Clusterpath An Algorithm for Clustering using Convex Fusion Penalties*, in "International Conference on Machine Learning (ICML)", 2011.

[11] J. LEZAMA, K. ALAHARI, J. SIVIC, I. LAPTEV. *Track to the Future: Spatio-temporal Video Segmentation with Long-range Motion Cues*, in "IEEE Conference on Computer Vision and Pattern Recognition", 2011.

[12] A. MISHRA, K. ALAHARI, C. V. JAWAHAR. *An MRF Model for Binarization of Natural Scene Text*, in "Proceedings of International Conference on Document Analysis and Recognition", 2011.

[13] K. RAJA, I. LAPTEV, P. PEREZ, L. OISEL. *Joint pose estimation and action recognition in image graphs*, in "International Conference on Image Processing", 2011.

[14] M. RODRIGUEZ, I. LAPTEV, J. SIVIC, J.-Y. AUDIBERT. *Density-aware person detection and tracking in crowds*, in "International Conference on Computer Vision", 2011.

[15] M. RODRIGUEZ, J. SIVIC, I. LAPTEV, J.-Y. AUDIBERT. *Data-driven Crowd Analysis*, in "International Conference on Computer Vision", 2011.

[16] B. C. RUSSELL, J. SIVIC, J. PONCE, H. DESSALES. *Automatic Alignment of Paintings and Photographs Depicting a 3D Scene*, in "3rd International IEEE Workshop on 3D Representation for Recognition (3dRR-11), with ICCV 2011", 2011.

[17] M. SCHMIDT, K. ALAHARI. *Generalized Fast Approximate Energy Minimization via Graph Cuts: Alpha-Expansion Beta-Shrink Moves*, in "UAI 2011 - 27th Conference on Uncertainty in Artificial Intelligence", Barcelona, Spain, July 2011, http://hal.inria.fr/inria-00617524/en.

[18] A. TORII, J. SIVIC, T. PAJDLA. *Visual localization by linear combination of image descriptors*, in "Proceedings of the 2nd IEEE Workshop on Mobile Vision, with ICCV 2011", 2011.

[19] O. WHYTE, J. SIVIC, A. ZISSERMAN. *Deblurring Shaken and Partially Saturated Images*, in "Proceedings of the IEEE Workshop on Color and Photometry in Computer Vision, with ICCV 2011", 2011.

[20] J. C. VAN GEMERT. *Exploiting Photographic Style for Category-Level Image Classification by Generalizing the Spatial Pyramid*, in "Intl. Conf. Multimedia Information Retrieval", 2011.

Scientific Books (or Scientific Book chapters)

[21] D. FORSYTH, J. PONCE. *Computer Vision: A Modern Approach. (Second edition)*, Pearson Education Inc., 2011.

Research Reports

[22] F. COUZINIE-DEVY, J. MAIRAL, F. BACH, J. PONCE. *Dictionary Learning for Deblurring and Digital Zoom*, INRIA, September 2011, http://hal.inria.fr/inria-00627402/en.