IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2012

Project-Team WILLOW

Models of visual object recognition and scene understanding

IN COLLABORATION WITH: Département d'Informatique de l'Ecole Normale Supérieure

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

Creation of the Project-Team: June 01, 2007 .

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 [Professor in the Engineering Department of the University of Oxford, and part-time professor at ENS, HdR]

PhD Students

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

Post-Doctoral Fellows

Karteek Alahari
Minsu Cho
Jian Sun
Visesh Chari

Administrative Assistant

Marine Meyer

Others

Petr Gronát
Anastasia Syromyatnikova
Senanayak Karri
Yves Ubelmann

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 and methodological research aimed at developing effective algorithms and architectures.

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: Piotr Bojanowski (Inria) and Vadim Kantorov (Inria). Minsu Cho, Visesh Chari and Jian Sun have been hired as post-docs. Alexei Efros (Professor, Carnegie Mellon University, USA) and René Vidal, (Associate Professor, Johns Hopkins University, USA) visited WILLOW in summer 2012. Five Willow PhD students have defended PhD in 2012: Y-Lan Boureau (currently postdoc at NYU), Olivier Duchenne (currently postdoc at CMU), Armand Joulin (currently postodc at Stanford Univ.), Muhammad Muneeb Ullah (currently faculty member at NUST-SEECS) and Oliver Whyte (currently at Microsoft).

2.2. Highlights of the Year

+ I. Laptev was awarded a Junior ERC Grant, starting in Jan 2013.
+ J. Sivic and I. Laptev (together with C. Schmid, Inria Grenoble) co-organized one week summer school on visual recognition and machine learning http://www.di.ens.fr/willow/events/cvml2012/. The school has attracted 181 participants from 34 countries.
+ J. Ponce became a senior member of the Institut Universitaire de France.
+ J. Ponce was awarded a US patent for the PMVS software developed in collaboration with Yasutaka Furukawa.

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. We have obtained a US patent 8,331,615 [1] for the corresponding software (PMVS, http://grail.cs.washington.edu/software/pmvs/) which is available under a GPL license and used for film production by ILM and Weta as well as by Google in Google Maps. It is also the basic technology used by Iconem, a start-up founded by Y. Ubelmann, a Willow collaborator.

for free for academics, and licensing negotiations with several companies are under way.

Our recent work (Russel *et al.*, 2011) has applied 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 archaeological sites. This direction is currently being continued in the PhD work of Mathieu Aubry. Our other current work outlined in detail in Section 6.1 is focused on (i) recovering indoor scene geometry from observations of person-object interactions video, (ii) visual place recognition in structured databases, where images are geotagged and organized in a graph, and (iii) developing a discriminative clustering approach able to discover geographically representative image elements from Google Street View imagery using only weak geographic supervision.

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 5.2. 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.

[1] The patent “Match, Expand, and Filter Technique for Multi-View Stereopsis,” was issued December 11, 2012 and assigned patent number 8,331,615.

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) strongly-supervised deformable part-based model for object recognition and localization, (ii) a MRF model for segmentation of text in natural scenes, and (iii) algorithms for multi-class cosegmentation using a novel energy-minimization approach based on the developed convex relaxation for weakly supervised classifiers.

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 (i) developing a geometrical model for removing image blur due to camera shake, (ii) preparing an online image deblurring demo, and (iii) developing new formulation for image deblurring cast as a multi-label energy minimization problem.

3.4. Human activity capture and classification

From a scientific point of view, visual action understanding is a computer vision problem that until recently has received little attention outside of extremely specific contexts such as surveillance or sports. Many of the 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. Within PhD of Piotr Bojanowski 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, scenes and primitive human actions pre-learned on related tasks. We also investigate highly-efficient methods for computing video features motivated by the need of processing very large and increasing amounts of video.

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

Actions of people are tightly coupled with their environments and surrounding objects. Moreover, object function can be learned and recognized from observations of person-object interactions in video and still images. Designing and learning models for person-object interactions, however, is a challenging task due to both (i) the huge variability in visual appearance and (ii) the lack of corresponding annotations. We address this problem by developing weakly-supervised techniques enabling learning interaction models from long-term observations of people in natural indoor video scenes such as obtained from time-lapse videos on YouTube. We also explore stereoscopic information in 3D movies to learn better models for people in video including person detection, segmentation, pose estimation, tracking and action recognition.

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. We are engaged in a project involving the archaeology laboratory at ENS and focusing on image-based artifact modeling and decorative pattern retrieval in Pompeii. This effort is part of the MSR-Inria project mentioned earlier and that will be discussed further later in this report. Application of our 3D reconstruction technology is now being explored in the field of cultural heritage and archeology by the start-up Iconem, founded by Y. Ubelmann, a Willow collaborator.

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. Patch-based Multi-view Stereo Software (PMVS)

PMVS is a multi-view stereo software that takes a set of images and camera parameters, then reconstructs 3D structure of an object or a scene visible in the images. Only rigid structure is reconstructed. The software outputs a set of oriented points instead of a polygonal (or a mesh) model, where both the 3D coordinate and the surface normal are estimated at each oriented point. The software and its documentation are available at http://www.di.ens.fr/pmvs/. The software is distributed under GPL. A US patent corresponding to this software "Match, Expand, and Filter Technique for Multi-View Stereopsis" was issued on December 11, 2012 and assigned patent number 8,331,615.

5.2. SPArse Modeling Software (SPAMS)

SPAMS v2.3 was released as open-source software in May 2012 (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/.

5.3. Local dense and sparse space-time features

This is a package with Linux binaries implementing extraction of local space-time features in video. We are preparing a new release of the code implementing highly-efficient video descriptors described in Section 6.4.5. Previous version of the package was released 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. Automatic Mining of Visual Architectural Elements

The code on automatic mining of visual architectural elements (v4.3) described in (Doersch *et al.* SIGGRAPH [6]) has been publicly released online in December 2012 (earlier version v3.0 was released in September 2012) at http://graphics.cs.cmu.edu/projects/whatMakesParis/paris_sigg_release.tar.gz.

5.5. Automatic Alignment of Paintings

The code for automatic alignment of paintings to a 3D model (Russell et al. 2011) was made publicly available in October 2012 at http://www.di.ens.fr/willow/research/paintingalignment/index.html.

5.6. Multi-Class Image Cosegmentation

This is a package of Matlab code implementing multi-class cosegmentation (Joulin *et al.* CVPR 2012 [13] and unsupervised discriminative clustering for image co-segmenting (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 September 2012 and is available at http://www.di.ens.fr/~joulin/code/DALCIM.zip.

5.7. Convex Relaxation of Weakly Supervised Models

This is a package of Matlab code implementing a general multi-class approach to weakly supervised classification described in (Joulin and Bach ICML 2012 [12]). The goal is to avoid local minima typically occuring expectation-maximization like algorithms and to optimize a cost function based on a convex relaxation of the soft-max loss. The package was last updated in September 2012 and is available at http://www.di.ens.fr/~joulin/code/ICML12_Joulin.zip.

5.8. Non-uniform Deblurring for Shaken Images

An online demo of non-uniform deblurring for shaken images implementing the algorithm described in [8] and (Whyte et al. CPCV 2011) was made available in 2012 at http://www.di.ens.fr/willow/research/deblurring/. The demo takes as an input an image uploaded by the user, automatically estimates the blur, and outputs the deblurred image.

6. New Results

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

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

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

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 pose estimation to extract functional and geometric constraints about the scene. 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 monocular time lapse sequences collected from YouTube and a dataset of still images of indoor scenes. We demonstrate that observing people performing different actions can significantly improve estimates of 3D scene geometry.

This work has been published in [11].

6.1.2. Learning and Calibrating Per-Location Classifiers for Visual Place Recognition

Participants: Petr Gronát, Josef Sivic, Guillaume Obozinski [Inria SIERRA], Tomáš Pajdla [CTU in Prague].

The aim of this work is to localize a query photograph by finding other images depicting the same place in a large geotagged image database. This is a challenging task due to changes in viewpoint, imaging conditions and the large size of the image database. The contribution of this work is two-fold. First, we cast the place recognition problem as a classification task and use the available geotags to train a classifier for each location in the database in a similar manner to per-exemplar SVMs in object recognition. Second, as only few positive training examples are available for each location, we propose a new approach to calibrate all the per-location SVM classifiers using *only* the negative examples. The calibration we propose relies on a significance measure essentially equivalent to the p-values classically used in statistical hypothesis testing. Experiments are performed on a database of 25,000 geotagged street view images of Pittsburgh and demonstrate improved place recognition accuracy of the proposed approach over the previous work. The problem addressed in this work is illustrated in Figure 1.

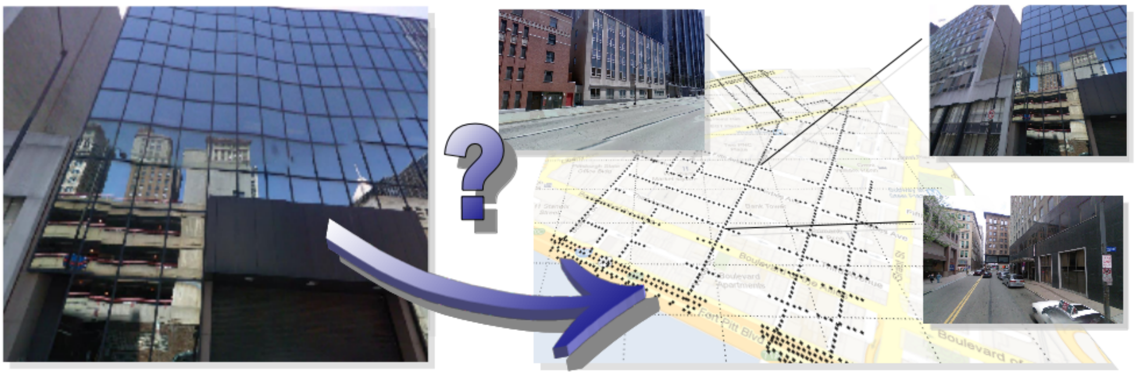

Figure 1. The goal of this work is to localize a query photograph (left) by finding other images of the same place in a large geotagged image database (right). We cast the problem as a classification task and learn a classifier for each location in the database. We develop a non-parametric procedure to calibrate the outputs of the large number of per-location classifiers without the need for additional positive training data.

This work has been submitted to CVPR 2013.

6.1.3. What Makes Paris Look like Paris?

Participants: Josef Sivic, Carl Doersch [CMU], Saurabh Singh [UIUC], Abhinav Gupta [CMU], Alexei Efros [CMU].

Given a large repository of geotagged imagery, we seek to automatically find visual elements, e.g. windows, balconies, and street signs, that are most distinctive for a certain geo-spatial area, for example the city of Paris. This is a tremendously difficult task as the visual features distinguishing architectural elements of different places can be very subtle. In addition, we face a hard search problem: given all possible patches in all images, which of them are both frequently occurring and geographically informative? To address these issues, we propose to use a discriminative clustering approach able to take into account the weak geographic supervision. We show that geographically representative image elements can be discovered automatically from Google Street View imagery in a discriminative manner. We demonstrate that these elements are visually interpretable and perceptually geo-informative. The discovered visual elements can also support a variety of computational geography tasks, such as mapping architectural correspondences and influences within and across cities, finding representative elements at different geo-spatial scales, and geographically-informed image retrieval. Example result is shown in Figure 2.

This work has been published in [6].

6.2. Category-level object and scene recognition

6.2.1. Task-Driven Dictionary Learning

Participants: Jean Ponce, Julien Mairal [Inria LEAR], 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

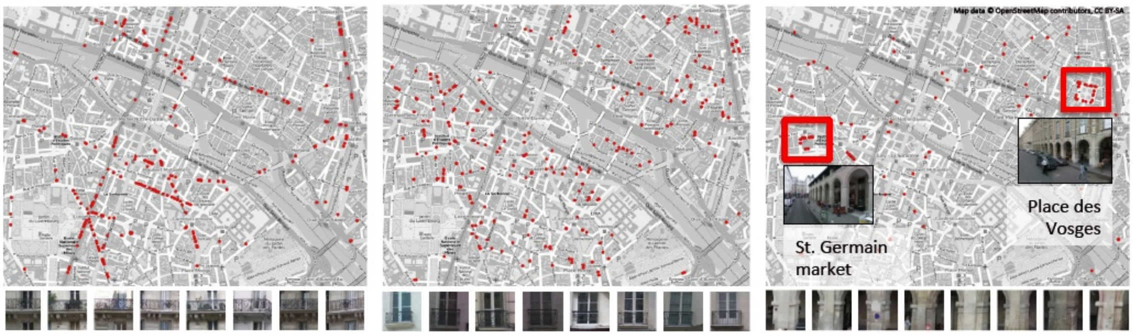

Figure 2. Examples of geographic patterns in Paris (shown as red dots on the maps) for three discovered visual elements (shown below each map). Balconies with cast-iron railings are concentrated on the main boulevards (left). Windows with railings mostly occur on smaller streets (middle). Arch supporting columns are concentrated on Place des Vosges and the St. Germain market (right).

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 been published in [7].

6.2.2. Object Detection Using Strongly-Supervised Deformable Part Models

Participants: Ivan Laptev, Hossein Azizpour [KTH].

Deformable part-based models achieve state-of-the-art performance for object detection, but rely on heuristic initialization during training due to the optimization of non-convex cost function. This work investigates limitations of such an initialization and extends earlier methods using additional supervision. We explore strong supervision in terms of annotated object parts and use it to (i) improve model initialization, (ii) optimize model structure, and (iii) handle partial occlusions. Our method is able to deal with sub-optimal and incomplete annotations of object parts and is shown to benefit from semi-supervised learning setups where part-level annotation is provided for a fraction of positive examples only. Experimental results are reported for the detection of six animal classes in PASCAL VOC 2007 and 2010 datasets. We demonstrate significant improvements in detection performance compared to the LSVM and the Poselet object detectors.

This work has been published in [9].

6.2.3. Multi-Class Cosegmentation

Participants: Armand Joulin, Jean Ponce, Francis Bach [Inria SIERRA].

Bottom-up, fully unsupervised segmentation remains a daunting challenge for computer vision. In the cosegmentation context, on the other hand, the availability of multiple images assumed to contain instances of the same object classes provides a weak form of supervision that can be exploited by discriminative approaches. Unfortunately, most existing algorithms are limited to a very small number of images and/or object classes (typically two of each). This work proposes a novel energy-minimization approach to cosegmentation that can handle multiple classes and a significantly larger number of images. The proposed cost function combines spectral- and discriminative-clustering terms, and it admits a probabilistic interpretation. It is optimized using an efficient EM method, initialized using a convex quadratic approximation of the energy.

Comparative experiments show that the proposed approach matches or improves the state of the art on several standard datasets.

This work has been published in [13].

6.2.4. A Convex Relaxation for Weakly Supervised Classifiers

Participants: Armand Joulin, Francis Bach [Inria SIERRA].

This work introduces a general multi-class approach to weakly supervised classification. Inferring the labels and learning the parameters of the model is usually done jointly through a block-coordinate descent algorithm such as expectation-maximization (EM), which may lead to local minima. To avoid this problem, we propose a cost function based on a convex relaxation of the soft-max loss. We then propose an algorithm specifically designed to efficiently solve the corresponding semidefinite program (SDP). Empirically, our method compares favorably to standard ones on different datasets for multiple instance learning and semi-supervised learning, as well as on clustering tasks.

This work has been published in [12].

6.2.5. Top-Down and Bottom-Up Cues for Scene Text Recognition

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 such text is a challenging problem, even more so than the recognition of scanned documents. In this work, we focus on the problem of recognizing text extracted from street images. 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 significant improvements in accuracies on two challenging public datasets, namely Street View Text (over 15%) and ICDAR 2003 (nearly 10%).

This work has been published in [15].

6.2.6. Scene Text Recognition using Higher Order Language Priors

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

The problem of recognizing text in images taken in the wild has gained significant attention from the computer vision community in recent years. Contrary to recognition of printed documents, recognizing scene text is a challenging problem. We focus on the problem of recognizing text extracted from natural scene images and the web. Significant attempts have been made to address this problem in the recent past. However, many of these works benefit from the availability of strong context, which naturally limits their applicability. In this work we present a framework that uses a higher order prior computed from an English dictionary to recognize a word, which may or may not be a part of the dictionary. We show experimental results on publicly available datasets. Furthermore, we introduce a large challenging word dataset with five thousand words to evaluate various steps of our method exhaustively.

The main contributions of this work are: (1) We present a framework, which incorporates higher order statistical language models to recognize words in an unconstrained manner (i.e. we overcome the need for restricted word lists, and instead use an English dictionary to compute the priors). (2) We achieve significant improvement (more than 20%) in word recognition accuracies without using a restricted word list. (3) We introduce a large word recognition dataset (at least 5 times larger than other public datasets) with character level annotation and benchmark it.

This work has been published in [14].

6.3. Image restoration, manipulation and enhancement

6.3.1. Non-Uniform Deblurring for Shaken Images

Participants: Josef Sivic, Andrew Zisserman, Jean Ponce, Oliver Whyte [Microsoft Redmond].

Photographs taken in low-light conditions are often blurry as a result of camera shake, i.e. a motion of the camera while its shutter is open. Most existing deblurring methods model the observed blurry image as the convolution of a sharp image with a uniform blur kernel. However, we show that blur from camera shake is in general mostly due to the 3D rotation of the camera, resulting in a blur that can be significantly non-uniform across the image. We propose a new parametrized geometric model of the blurring process in terms of the rotational motion of the camera during exposure. This model is able to capture non-uniform blur in an image due to camera shake using a single global descriptor, and can be substituted into existing deblurring algorithms with only small modifications. To demonstrate its effectiveness, we apply this model to two deblurring problems; first, the case where a single blurry image is available, for which we examine both an approximate marginalization approach and a maximum a posteriori approach, and second, the case where a sharp but noisy image of the scene is available in addition to the blurry image. We show that our approach makes it possible to model and remove a wider class of blurs than previous approaches, including uniform blur as a special case, and demonstrate its effectiveness with experiments on synthetic and real images.

This work has been published in [8]. An image deblurring demo, described in section 5.8, has been made available online.

6.3.2. Learning to Estimate and Remove Non-uniform Image Blur

Participants: Florent Couzinie-Devy, Jian Sun, Karteek Alahari, Jean Ponce.

This work addresses the problem of restoring images subjected to unknown and spatially varying blur caused by defocus or linear (say, horizontal) motion. The estimation of the global (non-uniform) image blur is cast as a multi-label energy minimization problem. The energy is the sum of unary terms corresponding to learned local blur estimators, and binary ones corresponding to blur smoothness. Its global minimum is found using Ishikawa's method by exploiting the natural order of discretized blur values for linear motions and defocus. Once the blur has been estimated, the image is restored using a robust (non-uniform) deblurring algorithm based on sparse regularization with global image statistics. The proposed algorithm outputs both a segmentation of the image into uniform-blur layers and an estimate of the corresponding sharp image. We present qualitative results on real images, and use synthetic data to quantitatively compare our approach to the publicly available implementation of Chakrabarti et al. 2010.

This work has been submitted to CVPR 2013.

6.4. Human activity capture and classification

6.4.1. Scene Semantics from Long-Term Observation of People

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

Our everyday objects support various tasks and can be used by people for different purposes. While object classification is a widely studied topic in computer vision, recognition of object function, i.e., what people can do with an object and how they do it, is rarely addressed. In this work we construct a functional object description with the aim to recognize objects by the way people interact with them. We describe scene objects (sofas, tables, chairs) by associated human poses and object appearance. Our model is learned discriminatively from automatically estimated body poses in many realistic scenes. In particular, we make use of time-lapse videos from YouTube providing a rich source of common human-object interactions and minimizing the effort of manual object annotation. We show how the models learned from human observations significantly improve object recognition and enable prediction of characteristic human poses in new scenes. Results are shown on a dataset of more than 400,000 frames obtained from 146 time-lapse videos of challenging and realistic indoor scenes. Some of the estimated human poses and results of pixel-wise scene segmentation are shown in Figure 3.

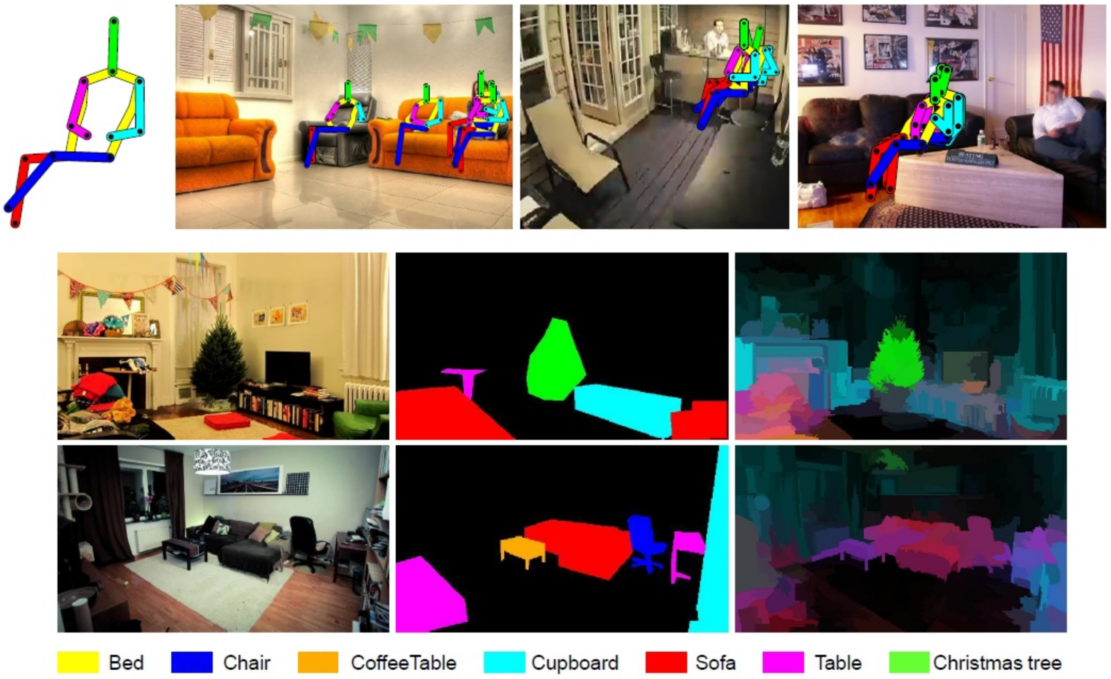

Figure 3. Top: Example of particular pose detections in three indoor scenes. Bottom: object segmentation illustrated by original images, ground truth segmentation, and automatic segmentation by our method shown in the left, middle and right columns respectively.

This work has been published in [10].

6.4.2. Analysis of Crowded Scenes in Video

Participants: Ivan Laptev, Josef Sivic, Mikel Rodriguez [MITRE].

In this work we first review the recent studies that have begun to address the various challenges associated with the analysis of crowded scenes. Next, we describe our two recent contributions to crowd analysis in video. First, we present a crowd analysis algorithm powered by prior probability distributions over behaviors that are learned on a large database of crowd videos gathered from the Internet. The proposed algorithm performs like state-of-the-art methods for tracking people having common crowd behaviors and outperforms the methods when the tracked individuals behave in an unusual way. Second, we address the problem of detecting and tracking a person in crowded video scenes. We formulate person detection as the optimization of a joint energy function combining crowd density estimation and the localization of individual people. The proposed methods are validated on a challenging video dataset of crowded scenes. Finally, the chapter concludes by describing ongoing and future research directions in crowd analysis.

This work is to appear in [17].

6.4.3. Actlets: A Novel Local Representation for Human Action Recognition in Video

Participants: Muhammad Muneeb Ullah, Ivan Laptev.

This work addresses the problem of human action recognition in realistic videos. We follow the recently successful local approaches and represent videos by means of local motion descriptors. To overcome the huge variability of human actions in motion and appearance, we propose a supervised approach to learn local motion descriptors – *actlets* – from a large pool of annotated video data. The main motivation behind our method is to construct action-characteristic representations of body joints undergoing specific motion patterns while learning invariance with respect to changes in camera views, lighting, human clothing, and other factors. We avoid the prohibitive cost of manual supervision and show how to learn actlets automatically from synthetic videos of avatars driven by the motion-capture data. We evaluate our method and show its significant improvement as well as its complementarity to existing techniques on the challenging UCF-sports and YouTube-actions datasets.

This work has been published in [16].

6.4.4. Layered Segmentation of People in Stereoscopic Movies

Participants: Karteek Alahari, Guillaume Seguin, Josef Sivic, Ivan Laptev.

In this work we seek to obtain a layered pixel-wise segmentation of multiple people in a stereoscopic video. This involves challenges such as dealing with unconstrained stereoscopic video, non-stationary cameras, complex indoor and outdoor dynamic scenes. The contributions of our work are three-fold: First, we develop a layered segmentation model incorporating person detections and pose estimates, as well as colour, motion, and stereo disparity cues. The model also explicitly represents depth ordering and occlusions of people. Second, we introduce a stereoscopic dataset with frames extracted from feature length movies “StreetDance 3D" and “Pina". In addition to realistic stereo image data, it contains nearly 700 annotated poses, 1200 annotated detections, and 400 pixel-wise segmentations of people. Third, we evaluate the benefits of stereo signal for person detection, pose estimation and segmentation in the new dataset. We demonstrate results on challenging realistic indoor and outdoor scenes depicting multiple people with frequent occlusions. Example result is shown in Figure 4.

This work has been submitted to CVPR 2013.

6.4.5. Highly-Efficient Video Features for Action Recognition and Counting

Participants: Vadim Kantorov, Ivan Laptev.

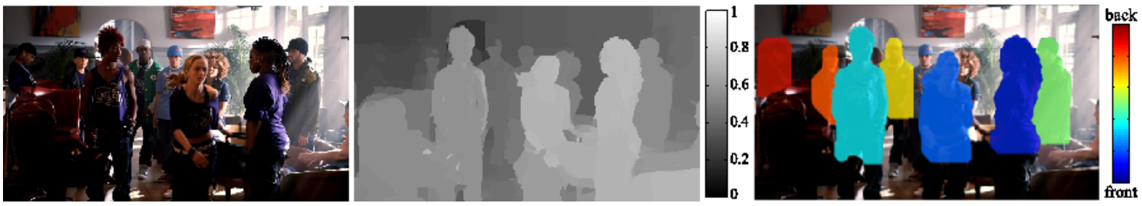

Figure 4. A sample frame extracted from the stereoscopic movie "StreetDance": From left to right – left image from the stereo pair, disparity map computed from the stereo pair, layered segmentation of the image into 7 people. The front to back ordering is shown as a colour map, where "blue" denotes front and "red" denotes back. The cost function associated with our model is initialized using person detections, and incorporates disparity, pose, colour and motion cues. Note that the result shows accurate segmentation boundaries and also a reliable layer ordering of people.

Local video features provide state-of-the-art performance for action recognition. While the accuracy of action recognition has been steadily improved over the recent years, the low speed of feature extraction remains to be a major bottleneck preventing current methods from addressing large-scale applications. In this work we demonstrate that local video features can be computed very efficiently by exploiting motion information readily-available from standard video compression schemes. We show experimentally that the use of sparse motion vectors provided by the video compression improves the speed of existing optical-flow based methods by two orders of magnitude while resulting in limited drops of recognition performance. Building on this representation, we next address the problem of event counting in video and present a method providing accurate counts of human actions and enabling to process 100 years of video on a modest computer cluster.

This work has been submitted to CVPR 2013.

7. Bilateral 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, 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. Google: Learning to annotate videos from movie scripts

Participants: Josef Sivic, Ivan Laptev, Jean Ponce.

The goal of this project is to automatically generate annotations of complex dynamic events in video. We wish to deal with events involving multiple people interacting with each other, objects and the scene, for example people at a party in a house. The goal is to generate structured annotations going beyond simple text tags. Examples include entire text sentences describing the video content as well as bounding boxes or segmentations spatially and temporally localizing the described objects and people in video. This is an extremely challenging task due to large intra-class variation of human actions. We propose to learn joint video and text representations enabling such annotation capabilities from feature length movies with coarsely aligned shooting scripts. Building on our previous work in this area, we aim to develop structured representations of video and associated text enabling to reason both spatially and temporally about scenes, objects and people as well as their interactions. Automatic understanding and interpretation of video content is a key-enabling factor for a range of practical applications such as content-aware advertising or search. Novel video and text representations are needed to enable breakthrough in this area.

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.

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.

8.3. International Initiatives

8.3.1. *IARPA FINDER Visual geo-localization (Inria)*

Participants: Josef Sivic, Petr Gronát.

Finder is an IARPA funded project aiming to develop technology to geo-localize images and videos that do not have geolocation tag. It is common today for even consumer-grade cameras to tag the images that they capture with the location of the image on the earth’s surface (“geolocation"). However, some imagery does not have a geolocation tag and it can be important to know the location of the camera, image, or objects in the scene. Finder aims to develop technology to automatically or semi-automatically geo-localize images and video that do not have the geolocation tag using reference data from many sources, including overhead and ground-based images, digital elevation data, existing well-understood image collections, surface geology, geography, and cultural information.

Partners: ObjectVideo, DigitalGlobe, CMU, Brown Univ., Cornell Univ., Univ. of Kentucky, GMU, Indiana Univ., and Washington Univ.

8.3.2. *Inria Associate Team VIP*

Participants: Ivan Laptev, Josef Sivic.

This project brings together three internationally recognized research groups with complementary expertise in human action recognition (Inria), qualitative and geometric scene interpretation (CMU) and large scale object recognition and human visual perception (MIT). The goal of VIP (Visual Interpretation of functional Properties) is to discover, model and learn functional properties of objects and scenes from image and video data.

Partners: Aude Oliva (MIT) and Alexei Efros (CMU). The project will be funded during 2012-2014.

8.4. International Research Visitors

8.4.1. *Visits of International Scientists*

Alexei Efros (Carnegie Mellon University) and René Vidal (Johns Hopkins University) have visited WIllow during summer 2012.

8.4.2. Visits to International Teams

Vincent Delaitre has visited the Robotics Institute, Carnegie Mellon University during November 2012 — January 2013, within the scope of Inria associate team VIP.

Armand Joulin has done a 3 months internship at Microsoft Research in Redmond, U.S.A.

9. Dissemination

9.1. Scientific Animation

+ Conference and workshop organization
  - K. Alahari, Co-organizer of the Workshop on Higher-Order Models and Global Constraints in Computer Vision in Computer Vision at ECCV 2012. https://sites.google.com/a/ttic.edu/eccv-2012-workshop-hipot/
  - I. Laptev, J. Sivic, Co-organizers of the First International Workshop on Action recognition and Pose Estimation in Still Images at ECCV 2012. http://vision.stanford.edu/apsi2012
  - A. Zisserman, Co-organizer of the PASCAL Visual Object Classes Challenge 2012 (VOC2012). http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/
  - A. Zisserman, Co-organizer of the PASCAL VOC 2012 workshop at ECCV 2012. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/workshop/index.html
+ 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).
  - IEEE Transactions on Pattern Analysis and Machine Intelligence (K. Alahari, co-guest editor of a special issue).
+ Area Chairs
  - European Conference on Computer Vision (ECCV), 2012 (I. Laptev, J. Ponce, J. Sivic, A. Zisserman).
  - IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013 (I. Laptev, J. Ponce, J. Sivic).
  - IEEE International Conference on Computer Vision (ICCV), 2013 (J. Sivic)
+ Program Committees
  - IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012 (M. Cho, I. Laptev, J. Sivic, A. Zisserman).
  - IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013 (K. Alahari, M. Cho, V. Delaitre, O. Duchenne, A. Joulin, V. Kantorov).
  - European Conference on Computer Vision (ECCV), 2012 (K. Alahari, M. Cho, V. Delaitre, O. Duchenne, A. Joulin).
  - Asian Conference on Computer Vision (ACCV), 2012 (K. Alahari).
  - Indian Conference on Computer Vision and Pattern Recognition (ICVGIP), 2012 (K. Alahari).
  - IEEE International Conference on Robotics and Automation (ICRA), 2012 (J. Sivic)
  - International Conference on Neural Information Processing Systems (NIPS), 2012 (K. Alahari, A. Joulin, I. Laptev, J. Sivic).
  - IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2012 (J. Sivic).
  - ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), 2012 (J. Sivic).
  - ACM International Conference on Multimedia Retrieval (ICMR), 2012 (J. Sivic)
  - International Conference on Artificial Intelligence and Statistics (AISTATS), 2012, 2013 (A. Joulin).
  - International Conference on Machine Learning (ICML), 2013 (A. Joulin).
  - A. Zisserman, reviewer for ERC grant applications.

+ Prizes and distinctions:
  - Sr. Member, Institut Universitaire de France (J. Ponce).
  - Inria Prime d'excellence scientifique (I. Laptev, J. Sivic).
  - ENS Prime d'excellence scientifique (J. Ponce).
+ Other:
  - PostDoc selection committee, Inria Paris, 2012 (I. Laptev)
  - Commission de Développement Technologique, Inria Paris, 2012 (J. Sivic)
  - Director, ENS Department of Computer Science (J. Ponce)

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence : J. Ponce, "Introduction to computer vision", L3, Ecole normale supérieure, 36h.

Licence : M. Pocchiola and J. Ponce, "Geometric bases of computer science", L3, Ecole normale supérieure, 36h.

Master : 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.

Doctorat : I. Laptev, J. Ponce, J. Sivic and A. Zisserman were speakers at the Inria Visual Recognition and Machine Learning Summer School http://www.di.ens.fr/willow/events/cvml2012/, Grenoble, 15h.

Doctorat : I. Laptev, speaker at the AERFAI Summer School on pattern recognition in multimodal human interaction, Vigo, Spain, 4h.

Doctorat : I. Laptev, speaker at the Human Activity and Vision Summer School, Inria Sophia-Antipolis, France, 2h.

9.2.2. Supervision

PhD : Oliver Whyte, "Removing Camera Shake Blur and Unwanted Occluders from Photographs" [5], École normale supérieure de Cachan, defended on March 15, 2012, J. Ponce, J. Sivic and A. Zisserman.

PhD : Y-Lan Boureau, "Learning Hierarchical Feature Extractors For Image Recognition" [1], New York University, September 2012, J. Ponce.

PhD : Muhammad Muneeb Ullah, "Supervised Statistical Representations for Human Action Recognition in Video" [4], Université de Rennes 1, defended on October 23, 2012, I. Laptev and P. Pérez.

PhD : Olivier Duchenne, "Non-rigid alignment methods for object recognition" [2], École normale supérieure, defended on November 29, 2012, J. Ponce.

PhD : Armand Joulin, "Convex optimization for cosegmentation" [3], École normale supérieure, defended on December 17, 2012, F. Bach and J. Ponce.

PhD in progress : Louise Benoît, started in 2009, J.Ponce.

PhD in progress : Piotr Bojanowski, "Learning to annotate dynamic video scenes", started in 2012, I. Laptev, J. Ponce, C. Schmid and J. Sivic.

PhD in progress : Florent Couzinié-Devy, started in 2009, J.Ponce.

PhD in progress : Vincent Delaitre, "Modeling and recognition of human-object interactions", started in 2010, I. Laptev and J. Sivic.

PhD in progress : Warith Harchaoui, "Modeling and alignment of human actions in video", started in 2011, I. Laptev, J. Ponce and J. Sivic.

PhD in progress : Vadim Kantorov, “Large-scale video mining and recognition”, started in 2012, I. Laptev.

PhD in progress : Guillaume Seguin, “Human action recognition using depth cues”, started in 2010, J. Sivic and I. Laptev.

PhD in progress : Marc Sturzel, started in 2008, J. Ponce, A. Zisserman.

9.2.3. Juries

+ PhD thesis committee:
  - Roland Angst, ETH Zurich, 2012 (J. Ponce).
  - Adrien Gaidon, Inria Rhone-Alpes, 2012 (I. Laptev).
  - Pyry Matikainen, CMU, 2012 (I. Laptev).
  - Olivier Duchenne, ENS Ulm, 2012 (J. Ponce).
  - Armand Joulin, ENS Ulm, 2012 (J. Ponce).
  - Gaurav Sharma, University of Caen, 2012 (J. Ponce).
  - Oliver Whyte, ENS Cachan, 2012 (J. Ponce, J. Sivic, A. Zisserman).
  - Muhammad Muneeb Ullah, ENS Cachan, 2012 (I. Laptev).
+ HDR thesis committee:
  - Florent Perronnin, Université Joseph Fourier. (J. Ponce).
+ Other:
  - "Jury prix Gilles Kahn" for best French PhD in computer science in November 2012 (J. Ponce).

9.3. Inria Visual Recognition and Machine Learning Summer School 2012

http://www.di.ens.fr/willow/events/cvml2012

I. Laptev 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 Inria Grenoble, attracted 181 participants from 34 countries (22% France / 48% Europe / 30% other countries (including Australia, Brazil, Canada, China, India, Iran, Israel, Japan, Malaysia, Mexico, Saudi Arabia, Singapore, South Korea, Turkey and USA)), and included Master students, PhD students as well as Post-docs and industrial participants. The summer school provided an overview of the state of the art in visual recognition and machine learning. Lectures were given by 14 speakers (6 USA, 1 UK, 1 Austria, 6 Inria / ENS), which included top international experts in the area of visual recognition (D. Forsyth, UIUC, USA; M. Hebert and A. Efros, CMU, USA; D. Ramanan, UC Irvine, USA, A. Torralba and A. Oliva, MIT, USA; A. Zisserman, Oxford, UK / WILLOW). 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 Fourth summer school in this series is currently in preparation for 2013 to be hosted by ENS Ulm and organized jointly by ENS and Inria Paris.

9.4. Invited presentations

- K. Alahari, Oxford Univ., Oxford, UK, Host: A. Zisserman, Sept. 2012.
- I. Laptev, 3rd AFCV Int. Workshop on Recent Trends in Computer Vision, Osaka, Japan, Jan. 2012.
- I. Laptev, 10th Workshop on Content Based Multimedia Indexing, Annecy, France, June 2012.
- I. Laptev, Oxford Univ., Oxford, UK, Host: A. Zisserman, Sept. 2012.
- I. Laptev, First Croatian Workshop on Computer Vision, Zagreb, Croatia, Sept. 2012.
- I. Laptev, 3rd IST Austria Symposium on Computer Vision and Machine Learning, Vienna, Austria, Oct. 2012.
- I. Laptev, Carnegie Mellon University, USA, Host: A. Efros, November 2012.
- J. Ponce, 2nd ACCV Workshop on e-Heritage, Daejeon, South Korea, Nov. 2012.
- J. Ponce, ETRI, Daejeon, South Korea, Nov. 2012.
- J. Ponce, “Let’s imagine the future”, Inria Rennes, France, Nov. 2012.
- J. Sivic, Google, USA, Hosts: H. Adam and H. Neven, August 2012
- J. Sivic, UC Berkeley, USA, Host: J. Malik, August 2012
- J. Sivic, Simon Fraser University, Canada, Host: G. Mori, August 2012
- J. Sivic, GdR ISIS workshop, Telecom ParisTech, Host: F. Lafarge, October 2012
- J. Sivic, First Int. Workshop on Visual Analysis and Geo-Localization of Large-Scale Imagery, ECCV 2012, October 2012.
- J. Sivic, Int. Workshop on Search Compupting, Brussels, Host: A. Joly, September 2012.
- J. Sivic, Carnegie Mellon University, USA, Host: A. Efros, December 2012.

9.5. Popularization

- The work “What Makes Paris Look like Paris” on the automatic mining of visual architectural elements (Doersch *et al.* SIGGRAPH [6]) has received broad press coverage including magazines The Wall Street Journal and NewScientist.
- The Pompeii project was featured in the special issue of "Cahiers Science & Vie", on "the lost worlds" ("les mondes perdus"), N°130, June 2012.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] Y.-L. Boureau. *Learning Hierarchical Feature Extractors For Image Recognition*, New York University, 2012, http://cs.nyu.edu/web/Research/Theses/boureau_y.pdf.

[2] O. Duchenne. *Non-rigid alignment methods for object recognition*, École normale supérieure, 2012, http://www.di.ens.fr/~duchenne/thesis.pdf.

[3] A. Joulin. *Convex optimization for cosegmentation*, École normale supérieure, 2012, http://www.di.ens.fr/~joulin/thesis.pdf.

[4] M. M. Ullah. *Supervised Statistical Representations for Human Action Recognition in Video*, Université de Rennes 1, 2012, http://www.di.ens.fr/willow/pdfscurrent/2012thesisUllah.pdf.

[5] O. WHYTE. *Removing Camera Shake Blur and Unwanted Occluders from Photographs*, Ecole Normale Supérieure de Cachan, 2012, http://www.di.ens.fr/willow/pdfscurrent/thesisWhyte.pdf.

Articles in International Peer-Reviewed Journals

[6] C. DOERSCH, S. SINGH, A. GUPTA, J. SIVIC, A. EFROS. *What Makes Paris Look like Paris?*, in "ACM Transactions on Graphics (SIGGRAPH)", 2012, vol. 31, n^{o} 4.

[7] J. MAIRAL, F. BACH, J. PONCE. *Task-Driven Dictionary Learning*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2012, vol. 34, n^{o} 4, p. 791–804, http://arxiv.org/abs/1009.5358.

[8] O. WHYTE, J. SIVIC, A. ZISSERMAN, J. PONCE. *Non-uniform Deblurring for Shaken Images*, in "International Journal of Computer Vision", 2012, vol. 98, n^{o} 2, p. 168–186.

International Conferences with Proceedings

[9] H. AZIZPOUR, I. LAPTEV. *Object detection using strongly-supervised deformable part models*, in "European Conference on Computer Vision", Florence, Italy, 2012.

[10] V. DELAITRE, D. FOUHEY, I. LAPTEV, J. SIVIC, A. EFROS, A. GUPTA. *Scene semantics from long-term observation of people*, in "European Conference on Computer Vision", Florence, Italy, 2012.

[11] D. FOUHEY, V. DELAITRE, A. EFROS, A. GUPTA, I. LAPTEV, J. SIVIC. *People Watching: Human Actions as a Cue for Single View Geometry*, in "European Conference on Computer Vision", Florence, Italy, 2012.

[12] A. JOULIN, F. BACH. *A convex relaxation for weakly supervised classifiers*, in "International Conference on Machine Learning", Edinburgh, United Kingdom, June 2012, 641, http://hal.inria.fr/hal-00717450.

[13] A. JOULIN, F. BACH, J. PONCE. *Multi-Class Cosegmentation*, in "IEEE Conference on Computer Vision and Pattern Recognition", Providence, United States, June 2012, 0109, http://hal.inria.fr/hal-00717448.

[14] A. MISHRA, K. ALAHARI, C. V. JAWAHAR. *Scene Text Recognition using Higher Order Language Priors*, in "British Machine Vision Conference", 2012.

[15] A. MISHRA, K. ALAHARI, C. V. JAWAHAR. *Top-Down and Bottom-Up Cues for Scene Text Recognition*, in "IEEE Conference on Computer Vision and Pattern Recognition", 2012.

[16] M. M. ULLAH, I. LAPTEV. *Actlets: A novel local representation for human action recognition in video*, in "IEEE International Conference on Image Processing", Orlando, Florida, USA, 2012.

Scientific Books (or Scientific Book chapters)

[17] M. RODRIGUEZ, J. SIVIC, I. LAPTEV. *Analysis of Crowded Scenes in Video*, in "Video analysis tools: applications in video surveillance", J.-Y. DUFOUR, P. MOUTTOU (editors), Hermes Science Publishing, 2012, In French, to appear.