IN PARTNERSHIP WITH:
**CNRS**

**Ecole normale supérieure de Paris**

# Activity Report 2013

# 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 interpretation**

# Table of contents

## Project-Team WILLOW

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

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

# 1. Members

**Research Scientists**

Ivan Laptev [Inria, Senior Researcher, HdR]
Josef Sivic [Inria, Researcher]

**Faculty Members**

Jean Ponce [Team leader, ENS Paris, Professor, HdR]
Andrew Zisserman [University of Oxford, Professor, HdR]

**Engineers**

Nicolas Maisonneuve [Inria]
Anastasia Syromyatnikova [Inria]

**PhD Students**

Mathieu Aubry [ENPC]
Louise Benoit [ENS Cachan]
Piotr Bojanowski [Inria]
Florent Couzinie-Devy [ENS Cachan]
Vincent Delaitre [ENS Lyon]
Warith Harchaoui [MSR-Inria]
Vadim Kantorov [Inria]
Maxime Oquab [Inria, from Jan 2014]
Rafael Sampaio de Rezende [Inria]
Guillaume Seguin [ENS Paris]
Marc Sturzel [EADS]
Tuan Hung Vu [Inria]

**Post-Doctoral Fellows**

Karteek Alahari [Inria, until Aug 2013]
Minsu Cho [Inria]
Visesh Chari [Inria]
Jian Sun [Inria]

**Visiting Scientists**

Petr Gronat [Phd student at Czech Technical University in Prague]
Alyosha Efros [Professor at UC Berkeley]
Aude Oliva [Principal Investigator at Massachusetts Institute of Technology]
John Canny [Professor at UC Berkeley]

**Administrative Assistant**

Marine Meyer [Inria]

# 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 three new Phd students: Maxime Oquab (MSR-Inria, starting Jan 2014), Rafael Sampaio de Rezende (Inria) and Tuang Hung VU (Inria). Aude Oliva (Principal Investigator, Massachusetts Institute of Technology, USA) and Alexei Efros (Professor, UC Berkeley, USA) visited Willow for 5 and 6 months, respectively. John Canny (Professor, UC Berkeley, USA) spent one week in Willow in November 2013 to begin a lasting collaboration with regular visits.

## 2.2. Highlights of the Year

- J. Sivic was awarded a Starting ERC Grant (2014-2018).
- J. Sivic, I. Laptev and J. Ponce (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/cvml2013/. The school has attracted 177 participants from 34 countries including Australia, Brazil, Canada, China, Japan, Korea, Russia, Singapore and the United States.

# 3. Research Program

## 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. We have also 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 (Russel *et al.*, 2011). Our current efforts in this area, outlined in detail in Section 6.1, are focused on: (i) developing new representations of 3D architectural sites for matching and retrieval, (ii) large-scale visual place recognition in structured image collections of urban environments, and (iii) continuing our theoretical study of multi-view camera geometry.

## 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), has focused on: (i) learning the appearance of objects and their parts in a weakly supervised manner, (ii) capturing the spatial layout of objects using the formalism of graph matching, (iii) developing models explicitly capturing the 3D structure of objects, and (iv) transferring mid-level image representations using convolutional neural networks.

## 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

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

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) investigating new geometrical models for removing image blur due to camera shake 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 the 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 and Platforms

## 5.1. SPArse Modeling Software (SPAMS)

SPAMS v2.4 was released as open-source software in May 2013 (v1.0 was released in September 2009, 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 ($\ell_1/\ell_2$, $\ell_1/\ell_\infty$, 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. 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 6.4.3. 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.

## 5.3. Automatic Mining of Visual Architectural Elements

The code on automatic mining of visual architectural elements (v4.5) described in (Doersch *et al.* SIGGRAPH 2012) has been publicly released online in January 2013 (earlier version v4.3 was released in December 2012 and v3.0 was released in September 2012) at http://graphics.cs.cmu.edu/projects/whatMakesParis/paris_sigg_release_v4.5.tar.gz.

## 5.4. Joint learning of actors and actions in video

This is a package of Matlab code implementing the multi-view face processing pipeline and joint learning of actors and actions in movies described in (Bojanowski *et al.* ICCV 2013 [2]. The package was last updated in December 2013 and is available at http://www.di.ens.fr/willow/research/actoraction/.

## 5.5. Visual Place Recognition with Repetitive Structures

Open-source release of the software package for visual localization in urban environments has been made publicly available. The software package implements newly developed method [9] for representing visual data containing repetitive structures (such as building facades or fences), which often occur in urban environments and present significant challenge for current image matching methods. The software is available at http://www.di.ens.fr/willow/research/repttile/download/repttile_demo_ver02.zip.

# 6. New Results

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

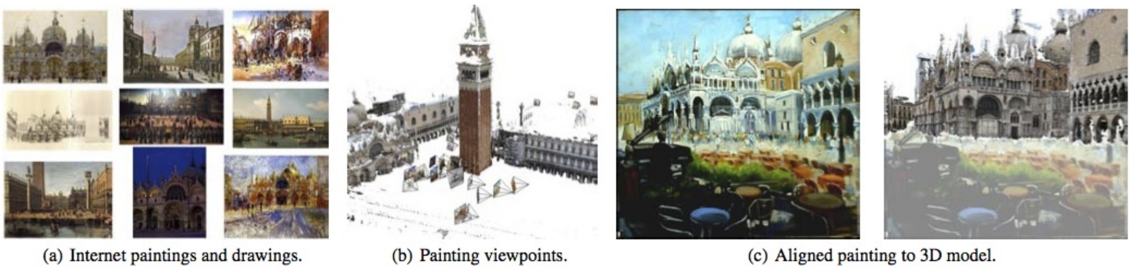
(a) Internet paintings and drawings. (b) Painting viewpoints. (c) Aligned painting to 3D model.

*Figure 1. Our system automatically aligns and recovers the viewpoint of paintings, drawings, and historical photographs to a 3D model of an architectural site.*

### *6.1.1. Painting-to-3D Model Alignment Via Discriminative Visual Elements*

**Participants:** Mathieu Aubry, Bryan Russell [Intel Labs], Josef Sivic.

In this work we describe a technique that can reliably align arbitrary 2D depictions of an architectural site, including drawings, paintings and historical photographs, with a 3D model of the site. This is a tremendously difficult task as the appearance and scene structure in the 2D depictions can be very different from the appearance and geometry of the 3D model, e.g., due to the specific rendering style, drawing error, age, lighting or change of seasons. In addition, we face a hard search problem: the number of possible alignments of the painting to a large 3D model, such as a partial reconstruction of a city, is huge. To address these issues, we develop a new compact representation of complex 3D scenes. The 3D model of the scene is represented by a small set of discriminative visual elements that are automatically learnt from rendered views. Similar to object detection, the set of visual elements, as well as the weights of individual features for each element, are learnt in a discriminative fashion. We show that the learnt visual elements are reliably matched in 2D depictions of the scene despite large variations in rendering style (e.g. watercolor, sketch, historical photograph) and structural changes (e.g. missing scene parts, large occluders) of the scene. We demonstrate an application of the proposed approach to automatic re-photography to find an approximate viewpoint of historical paintings and photographs with respect to a 3D model of the site. The proposed alignment procedure is validated via a human user study on a new database of paintings and sketches spanning several sites. The results demonstrate that our algorithm produces significantly better alignments than several baseline methods. This work has been accepted for publication to the ACM Transactions on Graphics (ACM ToG). The problem addressed in this work is illustrated in Figure 1 and example results are shown in figure 2. The pre-print is available online at [10].

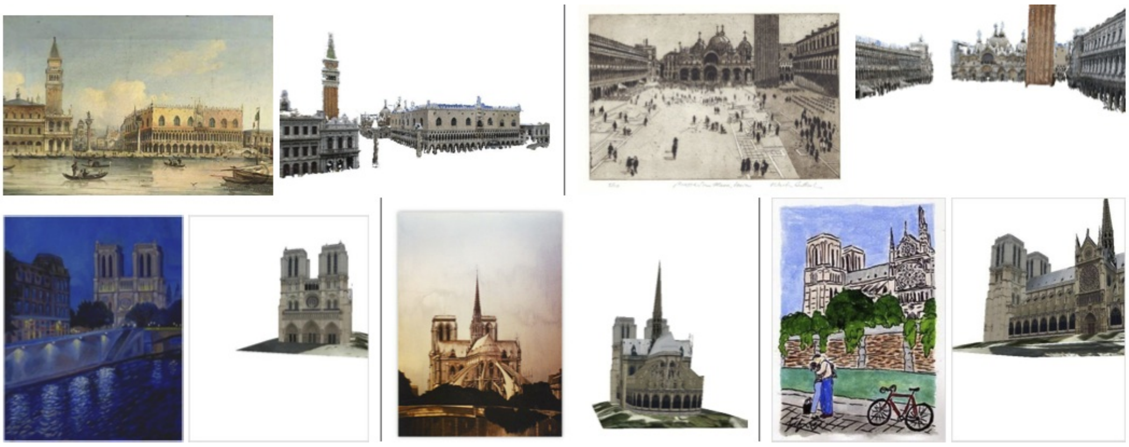

*Figure 2. Example alignments of non-photographic depictions to 3D models. Notice that we are able to align depictions rendered in different styles and having a variety of viewpoints with respect to the 3D models.*

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

**Participants:** Petr Gronat, Josef Sivic, Guillaume Obozinski [ENPC / 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. This work has been published at CVPR 2013 [6].

### *6.1.3. Visual Place Recognition with Repetitive Structures*

**Participants:** Akihiko Torii [Tokyo Institute of Technology], Josef Sivic, Tomáš Pajdla [CTU in Prague], Masatoshi Okutomi [Tokyo Institute of Technology].

Repeated structures such as building facades, fences or road markings often represent a significant challenge for place recognition. Repeated structures are notoriously hard for establishing correspondences using multi-view geometry. Even more importantly, they violate the feature independence assumed in the bag-of-visual-words representation which often leads to over-counting evidence and significant degradation of retrieval performance. In this work we show that repeated structures are not a nuisance but, when appropriately represented, they form an important distinguishing feature for many places. We describe a representation of repeated structures suitable for scalable retrieval. It is based on robust detection of repeated image structures and a simple modification of weights in the bag-of-visual-word model. Place recognition results are shown on datasets of street-level imagery from Pittsburgh and San Francisco demonstrating significant gains in recognition performance compared to the standard bag-of-visual-words baseline and more recently proposed burstiness weighting. This work has been published at CVPR 2013 [9].

### *6.1.4. Trinocular Geometry Revisited*

**Participants:** Jean Ponce, Martial Hebert [CMU].

When do the visual rays associated with triplets of point correspondences converge, that is, intersect in a common point? Classical models of trinocular geometry based on the fundamental matrices and trifocal tensor associated with the corresponding cameras only provide partial answers to this fundamental question, in large part because of underlying, but seldom explicit, general configuration assumptions. In this project, we use elementary tools from projective line geometry to provide necessary and sufficient geometric and analytical conditions for convergence in terms of transversals to triplets of visual rays, without any such assumptions. In turn, this yields a novel and simple minimal parameterization of trinocular geometry for cameras with non-collinear or collinear pinholes. This work has been submitted to CVPR 2014.

## 6.2. Category-level object and scene recognition

### *6.2.1. Learning Graphs to Match*

**Participants:** Minsu Cho, Karteek Alahari, Jean Ponce.

Many tasks in computer vision are formulated as graph matching problems. Despite the NP-hard nature of the problem, fast and accurate approximations have led to significant progress in a wide range of applications. Learning graph models from observed data, however, still remains a challenging issue. This work presents an effective scheme to parameterize a graph model, and learn its structural attributes for visual object matching. For this, we propose a graph representation with histogram-based attributes, and optimize them to increase the matching accuracy. Experimental evaluations on synthetic and real image datasets demonstrate the effectiveness of our approach, and show significant improvement in matching accuracy over graphs with pre-defined structures. The work is illustrated in Figure 3. This work has been published ICCV 2013 [3].

### *6.2.2. Finding Matches in a Haystack: A Max-Pooling Strategy for Graph Matching in the Presence of Outliers*

**Participants:** Minsu Cho, Olivier Duchenne [Intel], Jian Sun, Jean Ponce.

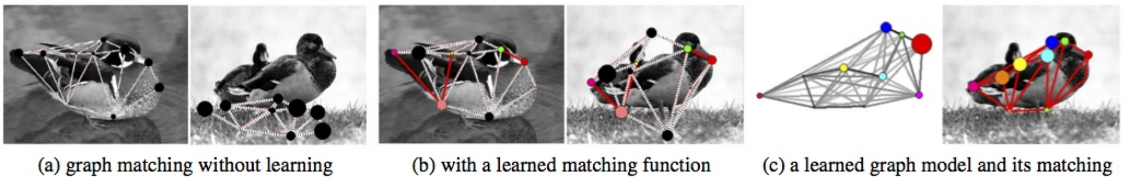
(a) graph matching without learning
(b) with a learned matching function
(c) a learned graph model and its matching

*Figure 3. Graph learning for matching. Our approach learns a graph model from labeled data to provide the best match to instances of a target class. It shows significant improvement over previous approaches for matching. (Best viewed in color.)*

A major challenge in real-world matching problems is to tolerate the numerous outliers arising in typical visual tasks. Variations in object appearance, shape, and structure within the same object class make it hard to distinguish inliers from outliers due to clutters. In this project, we propose a novel approach to graph matching, which is not only resilient to deformations but also remarkably tolerant to outliers. By adopting a max-pooling strategy within the graph matching framework, the proposed algorithm evaluates each candidate match using its most promising neighbors, and gradually propagates the corresponding scores to update the neighbors. As final output, it assigns a reliable score to each match together with its supporting neighbors, thus providing contextual information for further verification. We demonstrate the robustness and utility of our method with synthetic and real image experiments. This work has been submitted to CVPR 2014.

### 6.2.3. Decomposing Bag of Words Histograms

**Participants:** Ankit Gandhi [IIIT India], Karteek Alahari, C.v. Jawahar [IIIT India].

We aim to decompose a global histogram representation of an image into histograms of its associated objects and regions. This task is formulated as an optimization problem, given a set of linear classifiers, which can effectively discriminate the object categories present in the image. Our decomposition bypasses harder problems associated with accurately localizing and segmenting objects. We evaluate our method on a wide variety of composite histograms, and also compare it with MRF-based solutions. In addition to merely measuring the accuracy of decomposition, we also show the utility of the estimated object and background histograms for the task of image classification on the PASCAL VOC 2007 dataset. This work has been published at ICCV 2013 [5].

### 6.2.4. Image Retrieval using Textual Cues

**Participants:** Anand Mishra [IIIT India], Karteek Alahari, C.v. Jawahar [IIIT India].

We present an approach for the text-to-image retrieval problem based on textual content present in images. Given the recent developments in understanding text in images, an appealing approach to address this problem is to localize and recognize the text, and then query the database, as in a text retrieval problem. We show that such an approach, despite being based on state-of-the-art methods, is insufficient, and propose a method, where we do not rely on an exact localization and recognition pipeline. We take a query-driven search approach, where we find approximate locations of characters in the text query, and then impose spatial constraints to generate a ranked list of images in the database. The retrieval performance is evaluated on public scene text datasets as well as three large datasets, namely IIIT scene text retrieval, Sports-10K and TV series-1M, we introduce. This work has been published at ICCV 2013 [7].

### 6.2.5. Learning Discriminative Part Detectors for Image Classification and Cosegmentation

**Participants:** Jian Sun, Jean Ponce.

In this work, we address the problem of learning discriminative part detectors from image sets with category labels. We propose a novel latent SVM model regularized by group sparsity to learn these part detectors. Starting from a large set of initial parts, the group sparsity regularizer forces the model to jointly select and optimize a set of discriminative part detectors in a max-margin framework. We propose a stochastic version of a proximal algorithm to solve the corresponding optimization problem. We apply the proposed method to image classification and cosegmentation, and quantitative experiments with standard bench- marks show that it matches or improves upon the state of the art. This work has been published at CVPR 2013 [8].

### *6.2.6. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks*

**Participants:** Maxime Oquab, Leon Bottou [MSR New York], Ivan Laptev, Josef Sivic.

Convolutional neural networks (CNN) have recently shown outstanding image classification performance in the large-scale visual recognition challenge (ILSVRC2012). The success of CNNs is attributed to their ability to learn rich mid-level image representations as opposed to hand-designed low-level features used in other image classification methods. Learning CNNs, however, amounts to estimating millions of parameters and requires a very large number of annotated image samples. This property currently prevents application of CNNs to problems with limited training data. In this work we show how image representations learned with CNNs on large-scale annotated datasets can be efficiently transferred to other visual recognition tasks with limited amount of training data. We design a method to reuse layers trained on the ImageNet dataset to compute mid-level image representation for images in the PASCAL VOC dataset. We show that despite differences in image statistics and tasks in the two datasets, the transferred representation leads to significantly improved results for object and action classification, outperforming the current state of the art on Pascal VOC 2007 and 2012 datasets. We also show promising results for object and action localization. The pre-print of this work is available online [11]. Results are illustrated in Figure 4.

### *6.2.7. Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models*

**Participants:** Mathieu Aubry, Bryan Russell [Intel labs], Alyosha Efros [UC Berkeley], Josef Sivic.

We present an approach for the text-to-image retrieval problem based on textual content present in images. Given the recent developments in understanding text in images, an appealing approach to address this problem is to localize and recognize the text, and then query the database, as in a text retrieval problem. We show that such an approach, despite being based on state-of-the-art methods, is insufficient, and propose a method, where we do not rely on an exact localization and recognition pipeline. We take a query-driven search approach, where we find approximate locations of characters in the text query, and then impose spatial constraints to generate a ranked list of images in the database. The retrieval performance is evaluated on public scene text datasets as well as three large datasets, namely IIIT scene text retrieval, Sports-10K and TV series-1M, we introduce. This work has been submitted to CVPR 2014.

## 6.3. Image restoration, manipulation and enhancement

### *6.3.1. 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

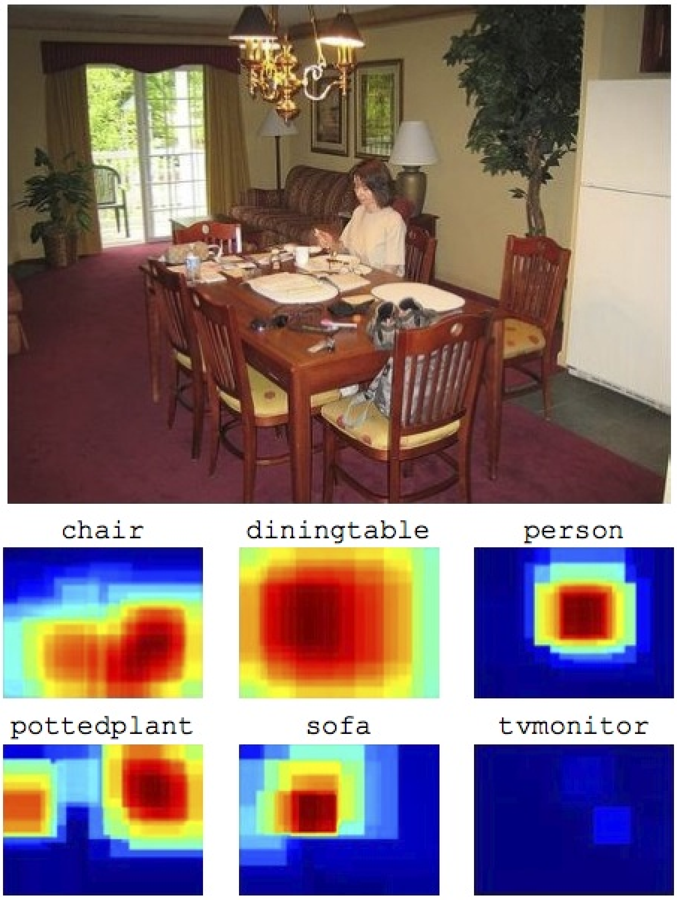

*Figure 4. Recognition and localization results of our method for a Pascal VOC test image. Output maps are shown for six object categories with the highest responses.*

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 published at CVPR 2013 [4] and example results are shown in figure 5.

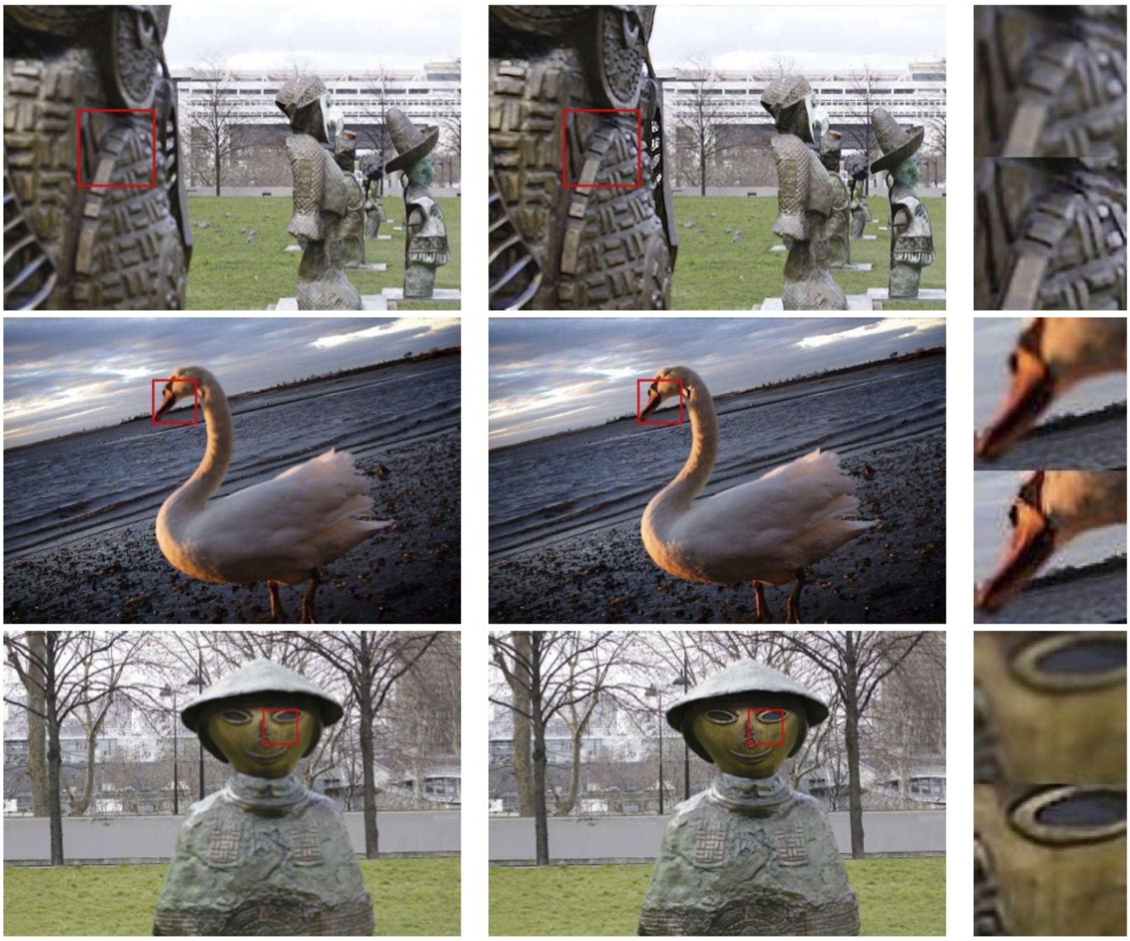

*Figure 5. Sample deblurring results on real images. From left to right: blurry image, deblurred image, close-up corresponding to the boxes shown in red. Note that our estimated deblurred image has more detail.*

### *6.3.2. Efficient, Blind, Spatially-Variant Deblurring for Shaken Images*

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

In this chapter we discuss modeling and removing spatially-variant blur from photographs. We describe a compact global parameterization of camera shake blur, based on the 3D rotation of the camera during the exposure. Our model uses three-parameter homographies to connect camera motion to image motion and, by assigning weights to a set of these homographies, can be seen as a generalization of the standard, spatially-invariant convolutional model of image blur. As such we show how existing algorithms, designed for spatially-invariant deblurring, can be “upgraded" in a straightforward manner to handle spatially-variant blur instead. We demonstrate this with algorithms working on real images, showing results for blind estimation of blur parameters from single images, followed by non-blind image restoration using these parameters. Finally, we introduce an efficient approximation to the global model, which significantly reduces the computational cost of modeling the spatially-variant blur. By approximating the blur as locally-uniform, we can take advantage of fast Fourier-domain convolution and deconvolution, reducing the time required for blind deblurring by an order of magnitude.

This work has been accepted for publication as a book chapter in the upcoming book “Motion Deblurring: Algorithms and Systems" to be published by Cambridge University Press in May 2014. [2] The demo implementing deblurring of images degraded by camera shake is available online at: http://www.di.ens.fr/willow/research/saturation/.

## 6.4. Human activity capture and classification

### *6.4.1. Layered Segmentation of People in Stereoscopic Movies*

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

In this work we seek to obtain a pixel-wise segmentation and pose estimation of multiple people in a stereoscopic video. This involves challenges such as dealing with unconstrained stereoscopic video, non-stationary cameras, and complex indoor and outdoor dynamic scenes. The contributions of our work are two-fold: First, we develop a segmentation model incorporating person detection, pose estimation, as well as colour, motion, and disparity cues. Our new model explicitly represents depth ordering and occlusion. Second, we introduce a stereoscopic dataset with frames extracted from feature-length movies “StreetDance 3D" and “Pina". The dataset contains 2727 realistic stereo pairs and includes annotation of human poses, person bounding boxes, and pixel-wise segmentations for hundreds of people. The dataset is composed of indoor and outdoor scenes depicting multiple people with frequent occlusions. We demonstrate results on our new challenging dataset, as well as on the H2view dataset from (Sheasby et al. ACCV 2012). This work has been published at ICCV 2013 [1].

### *6.4.2. Finding Actors and Actions in Movies*

**Participants:** Piotr Bojanowski, Francis Bach [Inria Sierra], Ivan Laptev, Jean Ponce, Cordelia Schmid [Inria Lear], Josef Sivic.

We address the problem of learning a joint model of actors and actions in movies using weak supervision provided by scripts. Specifically, we extract actor/action pairs from the script and use them as constraints in a discriminative clustering framework. The corresponding optimization problem is formulated as a quadratic program under linear constraints. People in video are represented by automatically extracted and tracked faces together with corresponding motion features. First, we apply the proposed framework to the task of learning names of characters in the movie and demonstrate significant improvements over previous methods used for this task. Second, we explore the joint actor/action constraint and show its advantage for weakly supervised action learning. We validate our method in the challenging setting of localizing and recognizing characters and their actions in feature length movies Casablanca and American Beauty. This work has been published at ICCV 2013 [2] and example results are shown in figure 6. The corresponding software has been also made publicly available (see the software section of this report).

### *6.4.3. Highly-Efficient Video Features for Action Recognition and Counting*

**Participants:** Vadim Kantorov, Ivan Laptev.

[2] http://www.cambridge.org/fr/academic/subjects/engineering/image-processing-and-machine-vision/motion-deblurring-algorithms-and-systems

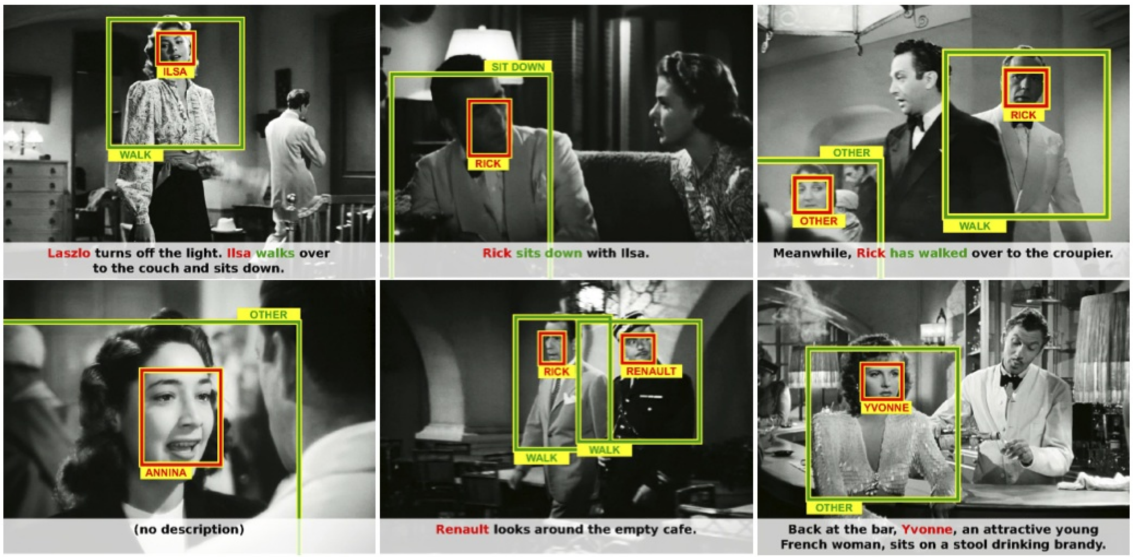
ILSA
WALK
Laszlo turns off the light. Ilsa walks over to the couch and sits down.
SIT DOWN
RICK
Rick sits down with Ilsa.
RICK
OTHER
OTHER
WALK
Meanwhile, Rick has walked over to the croupier.
OTHER
ANNINA
(no description)
RICK
RENAULT
WALK
WALK
Renault looks around the empty cafe.
YVONNE
OTHER
Back at the bar, Yvonne, an attractive young French woman, sits on a stool drinking brandy.

*Figure 6. Examples of automatically assigned names and actions in the movie Casablanca. Top row: Correct name and action assignments for tracks that have an actor/action constraint in the script. Bottom row: Correct name and action assignments for tracks that do not have a corresponding constraint in the script, but are still correctly classified. Note that even very infrequent characters are correctly classified (Annina and Yvonne). See more examples on the project web-page: http://www.di.ens.fr/willow/research/actoraction/*

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 2014.

# 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:** Leon Bottou [MSR], Ivan Laptev, Maxime Oquab, Jean Ponce, Josef Sivic, Cordelia Schmid [Inria Lear].

This collaborative project 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.

In October 2013 a new agreement has been signed for 2013-2016 with the research focus on automatic understanding of dynamic video content. Recent studies predict that by 2014 video will account for more than 90% of traffic on the Internet. Automatic understanding and interpretation of video content is a key enabling factor for a range of practical applications such as organizing and searching home videos or content aware video advertising. For example, interpreting videos of “making a birthday cake" or “planting a tree" could provide effective means for advertising products in local grocery stores or garden centers. The goal of this project is to perform fundamental computer science research in computer vision and machine learning in order to enhance the current capabilities to automatically understand, search and organize dynamic video content.

## 7.3. Google: Learning to annotate videos from movie scripts (Inria)

**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 (ANR): SEMAPOLIS*

**Participants:** Mathieu Aubry, Josef Sivic.

The goal of the SEMAPOLIS project is to develop advanced large-scale image analysis and learning techniques to semantize city images and produce semantized 3D reconstructions of urban environments, including proper rendering. Geometric 3D models of existing cities have a wide range of applications, such as navigation in virtual environments and realistic sceneries for video games and movies. A number of players (Google, Microsoft, Apple) have started to produce such data. However, the models feature only plain surfaces, textured from available pictures. This limits their use in urban studies and in the construction industry, excluding in practice applications to diagnosis and simulation. Besides, geometry and texturing are often wrong when there are invisible or discontinuous parts, e.g., with occluding foreground objects such as trees, cars or lampposts, which are pervasive in urban scenes. This project will go beyond the plain geometric models by producing semantized 3D models, i.e., models which are not bare surfaces but which identify architectural elements such as windows, walls, roofs, doors, etc. Semantic information is useful in a larger number of scenarios, including diagnosis and simulation for building renovation projects, accurate shadow impact taking into account actual window location, and more general urban planning and studies such as solar cell deployment. Another line of applications concerns improved virtual cities for navigation, with object-specific rendering, e.g., specular surfaces for windows. Models can also be made more compact, encoding object repetition (e.g., windows) rather than instances and replacing actual textures with more generic ones according to semantics; it allows cheap and fast transmission over low- bandwidth mobile phone networks, and efficient storage in GPS navigation devices.

This is a collaborative effort with LIGM / ENPC (R. Marlet), University of Caen (F. Jurie), Inria Sophia Antipolis (G. Drettakis) and Acute3D (R. Keriven).

## 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 labs: Mobile visual content analysis (Inria)*

**Participants:** Ivan Laptev, Josef Sivic.

The goal of this project within the European EIT-ICT activity is to mature developed technology towards real-world applications as well as transfer technology to industrial partners. Particular focus of this project is on computer vision technology for novel applications with wearable devices. The next generation mobile phones may not be in the pocket but worn by users as glasses continuously capturing audio-video data, providing visual feedback to the user and storing data for future access. Automatic answers to “Where did I leave my keys yesterday?" or “How did this place look like 100 years ago?" enabled by such devices could change our daily life while creating numerous new business opportunities. The output of this activity is new computer vision technology to enable a range of innovative mobile wearable applications.

This is a collaborative effort with S. Carlsson (KTH Stockholm) and J. Laaksonen (Aalto University).

### *8.2.3. European Research Council (ERC) Advanced Grant: “VideoWorld" - Jean Ponce*

**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.

### 8.2.4. *European Research Council (ERC) Starting Grant: “Activia" - Ivan Laptev*

**Participant:** Ivan Laptev.

WILLOW will be funded in part from 2013 to 2017 by the ERC Starting Grant "Activia" awarded to Ivan Laptev by the European Research Council.

Computer vision is concerned with the automated interpretation of images and video streams. Today’s research is (mostly) aimed at answering queries such as “Is this a picture of a dog?", “Is the person walking in this video?" (image and video categorisation) or sometimes “Find the dog in this photo" (object detection). While categorisation and detection are useful for many tasks, inferring correct class labels is not the final answer to visual recognition. The categories and locations of objects do not provide direct understanding of their function, i.e., how things work, what they can be used for, or how they can act and react. Neither do action categories provide direct understanding of subject’s intention, i.e., the purpose of his/her activity. Such an understanding, however, would be highly desirable to answer currently unsolvable queries such as “Am I in danger?" or “What can happen in this scene?". Answering such queries is the aim of this project.

The main challenge is to uncover the functional properties of objects and the purpose of actions by addressing visual recognition from a different and yet unexplored perspective. The major novelty of this proposal is to leverage observations of people, i.e., their actions and interactions to automatically learn the use, the purpose and the function of objects and scenes from visual data. This approach is timely as it builds upon two key recent technological advances: (a) the immense progress in visual object, scene and human action recognition achieved in the last ten years, and (b) the emergence of massive amounts of image and video data readily available for training visual models. My leading expertise in human action recognition and video understanding puts me in a strong position to realise this project. ACTIVIA addresses fundamental research issues in automated interpretation of dynamic visual scenes, but its results are expected to serve as a basis for ground-breaking technological advances in practical applications. The recognition of functional properties and intentions as explored in this project will directly support high-impact applications such as prediction and alert of abnormal events and automated personal assistance, which are likely to revolutionise today’s approaches to crime protection, hazard prevention, elderly care, and many others.

### 8.2.5. *European Research Council (ERC) Starting Grant: “Leap" - Josef Sivic*

**Participant:** Josef Sivic.

The contract is to be signed and will begin during 2014. WILLOW will be funded in part from 2014 to 2018 by the ERC Starting Grant "Leap" awarded to Josef Sivic by the European Research Council.

People constantly draw on past visual experiences to anticipate future events and better understand, navigate, and interact with their environment, for example, when seeing an angry dog or a quickly approaching car. Currently there is no artificial system with a similar level of visual analysis and prediction capabilities. LEAP is a first step in that direction, leveraging the emerging collective visual memory formed by the unprecedented amount of visual data available in public archives, on the Internet and from surveillance or personal cameras - a complex evolving net of dynamic scenes, distributed across many different data sources, and equipped with plentiful but noisy and incomplete metadata. The goal of this project is to analyze dynamic patterns in this shared visual experience in order (i) to find and quantify their trends; and (ii) learn to predict future events in dynamic scenes. With ever expanding computational resources and this extraordinary data, the main scientific challenge is now to invent new and powerful models adapted to its scale and its spatio-temporal, distributed and dynamic nature. To address this challenge, we will first design new models that generalize across different data sources, where scenes are captured under vastly different imaging conditions such as camera viewpoint, temporal sampling, illumination or resolution. Next, we will develop a framework for finding, describing and quantifying trends that involve measuring long-term changes in many related scenes. Finally, we will develop a methodology and tools for synthesizing complex future predictions from aligned past visual experiences. Our models will be automatically learnt from large-scale, distributed, and asynchronous visual data, coming from different sources and with different forms of readily-available but noisy and incomplete metadata such as text, speech, geotags, scene depth (stereo sensors), or gaze and body motion (wearable sensors). Breakthrough progress on these problems would have profound implications on our everyday lives as well as science and commerce, with safer cars that anticipate the behavior of pedestrians on streets; tools that help doctors monitor, diagnose and predict patients' health; and smart glasses that help people react in unfamiliar situations enabled by the advances from this project.

## 8.3. International Initiatives

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

**Participants:** Josef Sivic, Petr Gronat, Nicolas Maisonneuve.

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, UC Berkeley, 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.

### *8.3.3. Inria International Chair - Prof. John Canny (UC Berkeley)*

**Participants:** John Canny [UC Berkeley], Jean Ponce, Ivan Laptev, Josef Sivic.

Prof. John Canny (UC Berkeley) has been awarded the Inria International chair in 2013. He has visited Willow in November 2013 for a week to begin a lasting collaboration.

### *8.3.4. Inria CityLab initiative*

**Participants:** Josef Sivic, Jean Ponce, Ivan Laptev, Alyosha Efros [UC Berkeley].

Willow participates in the ongoing CityLab@Inria initiative (co-ordinated by V. Issarny), which aims to leverage Inria research results towards developing "smart cities" by enabling radically new ways of living in, regulating, operating and managing cities. The activity of Willow focuses on urban-scale quantitative visual analysis and is pursued in collaboration with A. Efros (UC Berkeley).

Currently, map-based street-level imagery, such as Google Street-view provides a comprehensive visual record of many cities worldwide. Additional visual sensors are likely to be wide-spread in near future: cameras will be built in most manufactured cars and (some) people will continuously capture their daily visual experience using wearable mobile devices such as Google Glass. All this data will provide large-scale, comprehensive and dynamically updated visual record of urban environments.

The goal of this project is to develop automatic data analytic tools for large-scale quantitative analysis of such dynamic visual data. The aim is to provide quantitative answers to questions like: What are the typical architectural elements (e.g., different types of windows or balconies) characterizing a visual style of a city district? What is their geo-spatial distribution (see figure 1)? How does the visual style of a geo-spatial area evolve over time? What are the boundaries between visually coherent areas in a city? Other types of interesting questions concern distribution of people and their activities: How do the number of people and their activities at particular places evolve during a day, over different seasons or years? Are there tourists sightseeing, urban dwellers shopping, elderly walking dogs, or children playing on the street? What are the major causes for bicycle accidents?

Break-through progress on these goals would open-up completely new ways smart cities are visualized, modeled, planned and simulated, taking into account large-scale dynamic visual input from a range of visual sensors (e.g., cameras on cars, visual data from citizens, or static surveillance cameras).

## 8.4. International Research Visitors

### *8.4.1. Visits of International Scientists*

Prof. Alexei Efros (UC Berkeley) has visited WIllow for six months in 2013. Aude Oliva (Principal investigator, Massachuesetts Institute of Technology) visited Willow for three months in 2013. Prof. John Canny (UC Berkeley) has visited Willow for a week in fall 2013 to begin a long term collaboration.

### *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 the Inria associate team VIP. Maxime Oquab has done a 3 months internship at Microsoft Research in New York City, U.S.A.

# 9. Dissemination

## 9.1. Scientific Animation

- Conference and workshop organization
  - Workshop co-organizer, ICCV'13 THUMOS: The First International Workshop on Action Recognition with a Large Number of Classes, Sydney, Australia, 2013 (Ivan Laptev).
- 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).
  - I. Laptev and J. Sivic co-edit a special issue on “Video representations for visual recognition" in the International Journal of Computer Vision.
  - J. Sivic co-edits a special issue on “Advances in Large-Scale Media Geo-Localization" in the International Journal of Computer Vision.
- Area Chairs
  - 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)
  - European Conference on Computer Vision (ECCV) 2014 (I. Laptev)

- Peer reviewing
  - IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013 (K. Alahari, M. Cho, V. Delaitre, V. Kantorov).
  - IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014 (K. Alahari, M. Cho, V. Delaitre, V. Kantorov, J. Sivic, I. Laptev).
  - IEEE International Conference on Robotics and Automation (ICRA), 2013 (J. Sivic).
  - International Conference on Neural Information Processing Systems (NIPS), 2013 (I. Laptev, J. Sivic).
  - ACM Symposium on User Interface Software and Technology (UIST) 2013 (I. Laptev).
  - ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), 2013 (J. Sivic).
  - ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH Asia), 2013 (J. Sivic).
  - I. Laptev, reviewer for ERC grant applications (2013).
- 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).

## 9.2. Teaching - Supervision - Juries

### *9.2.1. HdR*

- HdR: Ivan Laptev, Modeling and visual recognition of human actions and interactions, École normale supérieure

### *9.2.2. 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/ENS Visual Recognition and Machine Learning Summer School http://www.di.ens.fr/willow/events/cvml2013/, Paris, 15h.

Doctorat: I. Laptev, Tutorial on human action recognition at the Inria Visual Recognition and Machine Learning Summer School, Paris, July 2013.

Doctorat: I. Laptev, Short course on Visual object and action recognition, Samsung Electronics Polska, Warsaw, Poland, November 2013.

### *9.2.3. Supervision*

PhD in progress : Mathieu Aubry, "Visual recognition and retrieval of 3D objects and scenes", started in 2011, J. Sivic and D. Cremers (TU Munich).

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 : Maxime Oquab, “Learning to annotate dynamic scenes with convolutional neural networks”, started in Jan 2014, L. Bottou (MSR), I. Laptev and J. Sivic.

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.

PhD in progress : Tuang Hung VU, “Learning functional description of dynamic scenes”, started in 2013, I. Laptev.

### *9.2.4. Juries*

- PhD thesis committee:
  - Nicolas Ballas , Mines Paristech, 2013 (J. Ponce (rapporteur), J. Sivic).
  - Yanal Wazaefi (Aix Marseille), 2013 (J. Ponce).
  - Anton Andriyenko, TU Darmstadt, Germany, 2013 (I. Laptev (rapporteur)).
- HDR thesis committee:
  - Ivan Laptev, ENS Ulm, 2013 (J. Ponce).
- Other:
  - ICRA 2013 best computer vision paper award committee (J. Ponce).
  - Member of the PSL Research Council (J. Ponce).
  - Chair of the CMLA AERES evaluation committee at ENS Cachan, November 2013 (J. Ponce).
  - Member of the Mathematics and Systems Unit AERES evaluation committee at Mines ParisTech, January 2014 (J. Ponce).
  - Member of the Inria postdoc selection committee, 2012- (I. Laptev).
  - Member of Conseil de Laboratoire DI, ENS, 2013 (I. Laptev).
  - Member of Inria Commission de developpement technologique (CDT), 2012- (J. Sivic).

## 9.3. Inria/ENS Visual Recognition and Machine Learning Summer School 2013

http://www.di.ens.fr/willow/events/cvml2013

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, attracted 177 participants from 34 countries (28% France / 32% Europe / 40% other countries (including Australia, Brazil, Canada, China, India, Israel, Japan, Malaysia, 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 (5 USA, 1 UK, 1 Austria, 1 Germany, 6 France), which included top international experts in the area of visual recognition (J. Malik, UC Berkeley, USA; M. Hebert, CMU, USA; K. Grauman, UTexas, USA, A. Oliva, MIT, USA; Y. LeCun, NYU, USA; . 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.

## 9.4. Invited presentations

- I. Laptev, Invited talk at the 1st Workshop on Understanding Human Activities, Sydney, Australia, Dec. 2013.
- I. Laptev, Invited talk at the IEEEWorkshop on Decoding Subtle Cues from Social Interactions, Sydney, Australia, Dec. 2013.
- I. Laptev, Invited talk at the Int. Workshop on Action Similarity in Unconstrained Videos, Portland, Oregon, USA, June 2013.
- I. Laptev, Seminar at Technicolor, Paris, France, June 2013.
- I. Laptev, Seminar at TU Dadmstadt, Darmstadt, Germany, May 2013.
- I. Laptev, Seminar at Xerox Research Centre Europe, Grenoble, France, March 2013.
- I. Laptev, Seminar at University of Washington, Seattle, USA, Feb. 2013.
- J. Ponce, Invited talk at the 2013 Polaris Colloquium, Lille, January 2013
- J. Ponce, Invited talk at the The Google cultural center, June 2013
- J. Ponce, keynote speaker at the 2013 Annual Workshop of the Austrian Association for Pattern Recognition in Innsbruck, May 2013
- J. Ponce, 20th anniversary distinguished speaker at the Xerox Research Centre Europe in Grenoble, July 2013
- J. Sivic, ICCV AC workshop, Oxford, August 2013.
- J. Sivic, Seminar at Ecole Centrale, Paris, September 2013.
- J. Sivic, Invited talk at the Int. Workshop on Visual Analysis and Geo-Localization of Large-Scale Imagery, CVPR 2013, June 2013.
- J. Sivic, Invited talk at the 1st IEEE Workshop on Visual Analysis beyond Semantics, CVPR 2013, June 2013.
- J. Sivic, Invited talk at the IST Austria Symposium on Computer Vision and Machine Learning, Vienna, Austria, Sep. 2013.
- J. Sivic, Invited talk at the Workshop on Large-scale video search and mining, ICCV 2013.

## 9.5. Popularization

- J. Ponce participated in the iMatch meeting at Futur en Seine, June 2013
- J. Sivic participated in the CityLabs@Inria presentation at Futur en Seine, June 2013

# 10. Bibliography

## Publications of the year

### International Conferences with Proceedings

[1] K. Alahari, G. Seguin, J. Sivic, I. Laptev. *Pose Estimation and Segmentation of People in 3D Movies*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, 2013, http://hal.inria.fr/hal-00874884

[2] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, J. Sivic. *Finding Actors and Actions in Movies*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, 2013, http://hal.inria.fr/hal-00904991

[3] M. Cho, K. Alahari, J. Ponce. *Learning Graphs to Match*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, 2013, http://hal.inria.fr/hal-00875105

[4] F. Couzinie-Devy, J. Sun, K. Alahari, J. Ponce. *Learning to Estimate and Remove Non-uniform Image Blur*, in "CVPR 2013 - 26th IEEE Conference on Computer Vision and Pattern Recognition", Portland, United States, IEEE, 2013, pp. 1075-1082 [*DOI :* 10.1109/CVPR.2013.143], http://hal.inria.fr/hal-00818175

[5] A. Gandhi, K. Alahari, C. V. Jawahar. *Decomposing Bag of Words Histograms*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, 2013, http://hal.inria.fr/hal-00874895

[6] P. Gronat, G. Obozinski, J. Sivic, T. Pajdla. *Learning and calibrating per-location classifiers for visual place recognition*, in "CVPR 2013 - 26th IEEE Conference on Computer Vision and Pattern Recognition", Portland, United States, June 2013, http://hal.inria.fr/hal-00934332

[7] A. Mishra, K. Alahari, C. V. Jawahar. *Image Retrieval using Textual Cues*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, 2013, http://hal.inria.fr/hal-00875100

[8] J. Sun, J. Ponce. *Learning Discriminative Part Detectors for Image Classification and Cosegmentation*, in "ICCV 2013 - International conference on computer vision", Sydney, Australia, December 2013, http://hal.inria.fr/hal-00932380

[9] A. Torii, J. Sivic, T. Pajdla, M. Okutomi. *Visual Place Recognition with Repetitive Structures*, in "CVPR 2013 - 26th IEEE Conference on Computer Vision and Pattern Recognition", Portland, United States, IEEE, 2013, http://hal.inria.fr/hal-00934288

## Other Publications

[10] M. Aubry, B. Russell, J. Sivic. , *Painting-to-3D Model Alignment Via Discriminative Visual Elements*, September 2013, http://hal.inria.fr/hal-00863615

[11] M. Oquab, L. Bottou, I. Laptev, J. Sivic. , *Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks*, November 2013, http://hal.inria.fr/hal-00911179