IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2014

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, HdR]

Faculty Member
Jean Ponce [Team leader, ENS Paris, Professor]

Engineers
Cedric Doucet [Inria SED, part-time]
Petr Gronat [Inria]
Anastasia Syromyatnikova [Inria]

PhD Students
Mathieu Aubry [ENPC]
Louise Benoit [ENS Cachan, until Aug 2014]
Piotr Bojanowski [Inria]
Guilhem Chéron [Inria]
Florent Couzinié-Devy [ENS]
Théophile Dalens [Inria]
Vincent Delaitre [ENS Leon]
Warith Harchaoui [Inria, until May 2014]
Vadim Kantorov [Inria]
Maxime Oquab [Inria]
Rafael Sampaio de Rezende [Inria]
Guillaume Séguin [ENS Paris]
Matthew Trager [Inria]
Tuan Hung Vu [Inria]

Post-Doctoral Fellows
Relja Arandjelovic [Inria, from Oct 2014]
Minsu Cho [Inria]
Visesh Chari [Inria, until Nov 2014]
Bumsub Ham [Inria, from May 2014]
Suha Kwak [Inria, from Apr 2014]
Jian Sun [Inria, until Aug 2014]

Visiting Scientists
John Canny [Professor at UC Berkeley, Inria International Chair]
Alyosha Efros [Professor at UC Berkeley]

Administrative Assistants
David Dinis [Inria, from May 2014]
Marine Meyer [Inria, until May 2014]

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 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: Guilhem Chéron (MSR-Inria), Théophile Dalens (Inria) and Matthew Trager (Inria). Alexei Efros (Professor, UC Berkeley, USA) visited Willow during May-June. John Canny (Professor, UC Berkeley, USA) spent three month in Willow within the framework of Inria's International Chair program.

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 6.2, are focused on: (i) developing new representations of 3D architectural sites for matching and retrieval, (ii) modeling and recognition of objects in complex scenes using underlying 3D object models, 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 6.3, has focused on: (i) capturing the spatial layout of objects using the formalism of graph matching, (ii) transferring mid-level image representations using convolutional neural networks, and (iii) learning the appearance of objects and their parts in a weakly supervised manner.

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 6.4, has focused on (i) image editing using accelerated local Laplacian filters and (ii) developing new formulation for image deblurring cast as a deep learning 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

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

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 6.5.

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.

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. 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.

5. New Software and Platforms

5.1. SPArse Modeling Software (SPAMS)

SPAMS v2.5 was released as open-source software in May 2014 (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 (ℓ_1/ℓ_2, ℓ_1/ℓ_∞, sparse group lasso, tree-structured regularization, structured sparsity with overlapping groups,...).

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

5.2. Efficient video descriptors for action recognition

This package contains source code for highly-efficient extraction of local space-time video descriptors for action recognition. The accuracy of descriptors measured at standard benchmarks for action recognition is comparable to the state-of-the-art dense trajectory features, while being more than 100 times faster on standard CPU. The previous version of this code was evaluated in our recent work [12]. The package is available from http://www.di.ens.fr/~laptev/download/fastvideofeat-1.0.zip. Earlier version of our space-time video features is available at http://www.di.ens.fr/~laptev/download/stip-2.0-linux.zip.

5.3. Weakly Supervised Action Labeling in Videos Under Ordering Constraints

This is a package of Matlab code implementing weakly-supervised learning of actions from input videos and corresponding sequences of action labels. The code finds optimal alignment of action labels and video intervals during training. Along the optimization, the method trains corresponding action model. The package is available at http://www.di.ens.fr/willow/research/actionordering/. The method corresponding to this code package has been described and evaluated in Bojanowski *et al.* ECCV 2014 [10].

5.4. Visual Place Recognition with Repetitive Structures

Open-source release of the software package for visual localization in urban environments has been made publicly available in May 2014. The software package implements the method [A. Torii et al., CVPR 2013] 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. This is an extended version that includes geometric verification. The original version was released in 2013. The software is available at http://www.di.ens.fr/willow/research/repttile/download/repttile_demo_ver03.zip.

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

Open-source release of the software package for 2D-3D category-level alignment has been made publicly available. The software package implements newly developed method [9] for category-level recognition that not only outputs the bounding box of the object but predicts an approximate 3D model aligned with the input image. The software is available at http://www.di.ens.fr/willow/research/seeing3Dchairs/.

5.6. Painting-to-3D Model Alignment Via Discriminative Visual Elements

Open-source release of the software package for alignment of 3D models to historical and non-photographic depictions has been made publicly available. The software package implements the method of [9] for alignment of 3D models to input historical and non-photographic depictions such as paintings, drawings or engravings, where standard local feature-based method fail. The software is available at http://www.di.ens.fr/willow/research/painting_to_3d/.

5.7. Painting recognition from wearable cameras

Open-source release of the software package for painting recognition from wearable cameras has been made publicly available. The software implements a method described in [20] that recognizes 2D paintings on a wearable Google Glass device, for example, for a virtual museum guide application. The software runs directly on Google Glass without sending images to external servers for processing and recognizes a query painting in a database of 100 paintings in one second. The report and software are publicly available at http://www.di.ens.fr/willow/research/glasspainting/.

5.8. Learning and transferring mid-level image representations using convolutional neural networks

The first version of the open source software package for convolutional neural networks [13] has been released online. The software package is based on the cuda-convnet implementation of convolutional neural networks and includes a pre-trained convolutional neural network that can be applied to visual object classification as in the Pascal VOC 2012 set-up, where it achieves state-of-the-art single network results. The package also includes functions for visualization of object localization. The software is publicly available at http://www.di.ens.fr/willow/research/cnn/code/voc12-cvpr-reproduce.tar.

6. New Results

6.1. Highlights of the Year

- J. Sivic started ERC project LEAP (2014-2018).
- J. Sivic serves as a Program Chair for International Conference on Computer Vision, Santiago, Chile, 2015

6.2. 3D object and scene modeling, analysis, and retrieval

6.2.1. Painting-to-3D Model Alignment Via Discriminative Visual Elements

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

(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.

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 published at ACM Transactions on Graphics 2014 [3] and its extension has appeared at RFIA 2014 [17]. The problem addressed in this work is illustrated in Figure 1 and example results are shown in Figure 2.

6.2.2. 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.

This work poses object category detection in images as a type of 2D-to-3D alignment problem, utilizing the large quantities of 3D CAD models that have been made publicly available online. Using the "chair" class as a running example, we propose an exemplar-based 3D category representation, which can explicitly model chairs of different styles as well as the large variation in viewpoint. We develop an approach to establish part-based correspondences between 3D CAD models and real photographs. This is achieved by (i) representing each 3D model using a set of view-dependent mid-level visual elements learned from synthesized views in a discriminative fashion, (ii) carefully calibrating the individual element detectors on a common dataset of negative images, and (iii) matching visual elements to the test image allowing for small mutual deformations but preserving the viewpoint and style constraints. We demonstrate the ability of our system to align 3D models with 2D objects in the challenging PASCAL VOC images, which depict a wide variety of chairs in complex scenes. This work has been published at CVPR 2014 [9].

6.2.3. Anisotropic Laplace-Beltrami Operators for Shape Analysis

Participants: Mathieu Andreux [TUM], Emanuele Rodola [TUM], Mathieu Aubry, Daniel Cremers [TUM].

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.

This work introduces an anisotropic Laplace-Beltrami operator for shape analysis. While keeping useful properties of the standard Laplace-Beltrami operator, it introduces variability in the directions of principal curvature, giving rise to a more intuitive and semantically meaningful diffusion process. Although the benefits of anisotropic diffusion have already been noted in the area of mesh processing (e.g. surface regularization), focusing on the Laplacian itself, rather than on the diffusion process it induces, opens the possibility to effectively replace the omnipresent Laplace-Beltrami operator in many shape analysis methods. After providing a mathematical formulation and analysis of this new operator, we derive a practical implementation on discrete meshes. Further, we demonstrate the effectiveness of our new operator when employed in conjunction with different methods for shape segmentation and matching. This work has been published at the Sixth Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment (NORDIA) 2014 [8].

6.2.4. 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 published at CVPR 2014 [15].

6.2.5. On Image Contours of Projective Shapes

Participants: Jean Ponce, Martial Hebert [CMU].

This work revisits classical properties of the outlines of solid shapes bounded by smooth surfaces, and shows that they can be established in a purely projective setting, without appealing to Euclidean measurements such as normals or curvatures. In particular, we give new synthetic proofs of Koenderink's famous theorem on convexities and concavities of the image contour, and of the fact that the rim turns in the same direction as the viewpoint in the tangent plane at a convex point, and in the opposite direction at a hyperbolic point. This suggests that projective geometry should not be viewed merely as an analytical device for linearizing

calculations (its main role in structure from motion), but as the proper framework for studying the relation between solid shape and its perspective projections. Unlike previous work in this area, the proposed approach does not require an oriented setting, nor does it rely on any choice of coordinate system or analytical considerations. This work has been published at ECCV 2014 [14].

6.3. Category-level object and scene recognition

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

Participants: Minsu Cho, Jian Sun, Olivier Duchenne, Jean Ponce.

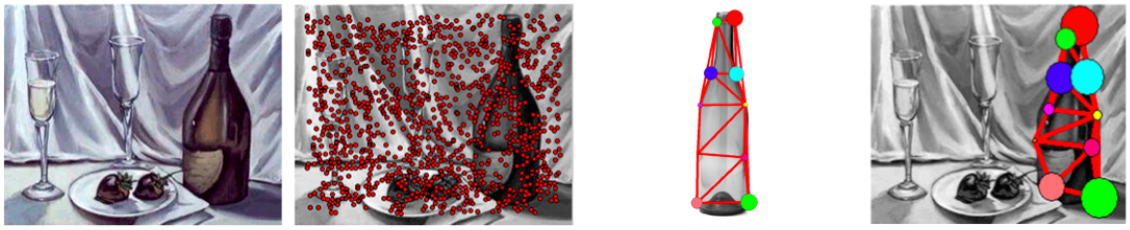

Figure 3. Feature matching in the presence of outliers. (a) In real-world scenes, background clutter often produces numerous outlier features, making it hard to find correspondences. (b) We address the issue with a max-pooling approach to graph matching. The proposed method is not only resilient to deformations but also remarkably tolerant to outliers. Each node on the left image corresponds to one with the same color on the right image, where bigger nodes represent more similar nodes. (Best viewed in color.)

A major challenge in real-world feature 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 harder to distinguish inliers from outliers due to clutters. In this work, we propose a max-pooling approach to graph matching, which is not only resilient to deformations but also remarkably tolerant to outliers. 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 published at CVPR 2014 [11]. The proposed method and its qualitative results are illustrated in Figure 3.

6.3.2. Unsupervised Object Discovery and Localization in the Wild: Part-based Matching with Bottom-up Region Proposals

Participants: Minsu Cho, Suha Kwak, Cordelia Schmid [Inria Lear], Jean Ponce.

This work addresses unsupervised discovery and localization of dominant objects from a noisy image collection of multiple object classes. The setting of this problem is fully unsupervised, without even image-level annotations or any assumption of a single dominant class. This is significantly more general than typical colocalization, cosegmentation, or weakly-supervised localization tasks. We tackle the discovery and localization problem using a part-based matching approach: We use off-the-shelf region proposals to form a set of candidate bounding boxes for objects and object parts. These regions are efficiently matched across images using a probabilistic Hough transform that evaluates the confidence in each candidate region

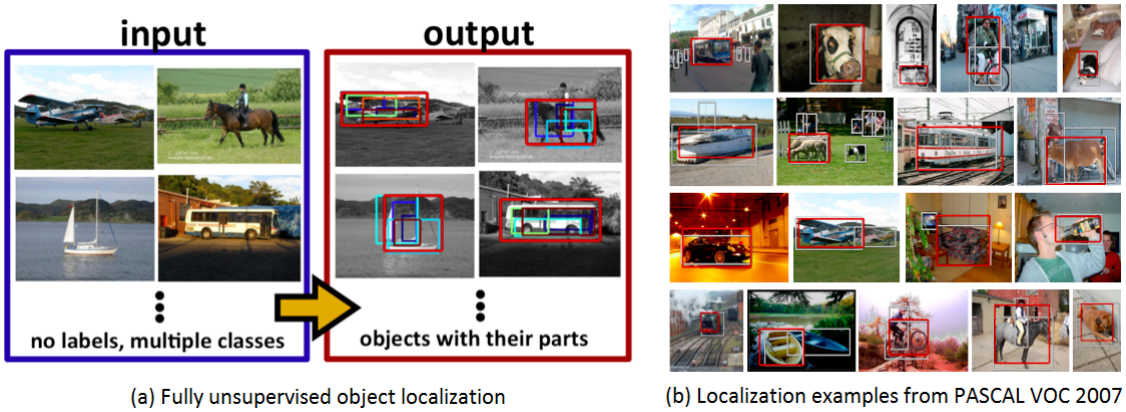

Figure 4. Unsupervised object discovery in the wild. (a) We tackle object localization in an unsupervised scenario without any types of annotations, where a given image collection may contain multiple dominant object classes and even outlier images. The proposed method discovers object instances (red bounding boxes) with their distinctive parts (smaller boxes). (b) Examples of localization on mixed-class PASCAL VOC 2007 train/val datasets are shown. Red boxes represent localized objects while white boxes are ground truth annotations. (Best viewed in color.)

considering both appearance similarity and spatial consistency. Dominant objects are discovered and localized by comparing the scores of candidate regions and selecting those that stand out over other regions containing them. Extensive experimental evaluations on standard benchmarks demonstrate that the proposed approach significantly outperforms the current state of the art in colocalization, and achieves robust object discovery in challenging mixed-class datasets. This work has been submitted to CVPR 2015 [22]. The proposed method and its qualitative results are illustrated in Figure 4.

6.3.3. *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. This work has been published at CVPR 2014 [13].

6.3.4. *Weakly supervised object recognition with convolutional neural networks*

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

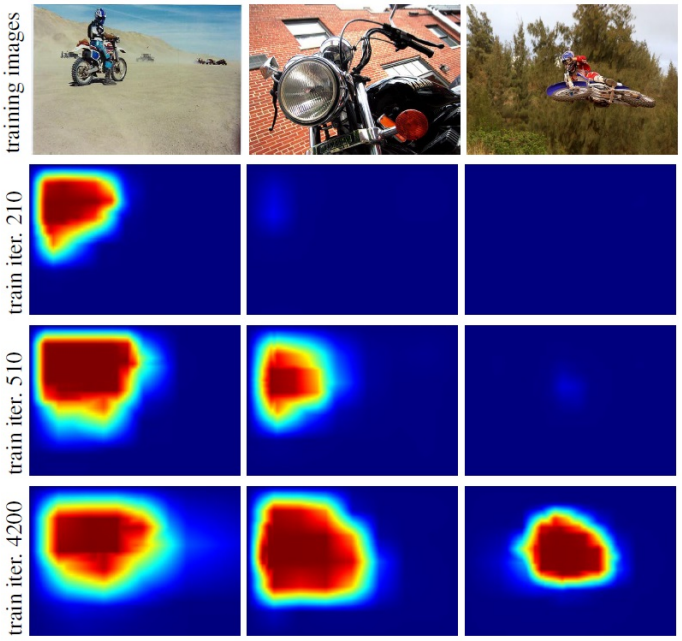

Figure 5. Evolution of localization score maps for the motorbike class over iterations of our weakly-supervised CNN training. Note that locations of objects with more usual appearance are discovered earlier during training.

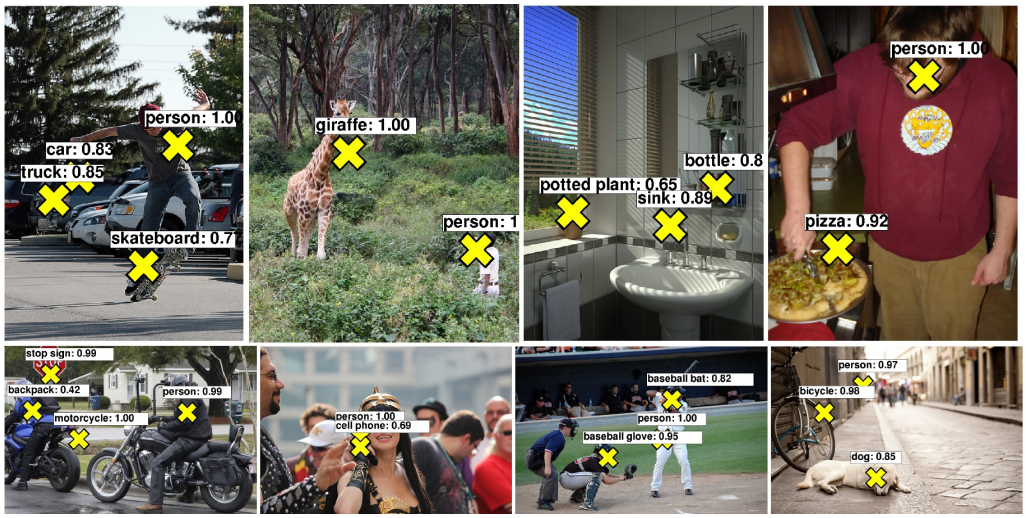

Figure 6. Example location predictions for images from the Microsoft COCO validation set obtained by our weakly-supervised method. Note that our method does not use object locations at training time, yet can predict locations of objects in test images (yellow crosses). The method outputs the most confident location for most confident object classes.

Successful methods for visual object recognition typically rely on training datasets containing lots of richly annotated images. Detailed image annotation, e.g. by object bounding boxes, however, is both expensive and often subjective. We describe a weakly supervised convolutional neural network (CNN) for object classification that relies only on image-level labels, yet can learn from cluttered scenes containing multiple objects (see Figure 5). We quantify its object classification and object location prediction performance on the Pascal VOC 2012 (20 object classes) and the much larger Microsoft COCO (80 object classes) datasets. We find that the network (i) outputs accurate image-level labels, (ii) predicts approximate locations (but not extents) of objects, and (iii) performs comparably to its fully-supervised counterparts using object bounding box annotation for training. This work has been submitted to CVPR 2015 [23]. Illustration of localization results by our method in Microsoft COCO dataset is illustrated in Figure 6.

6.3.5. Learning Dictionary of Discriminative Part Detectors for Image Categorization and Cosegmentation

Participants: Jian Sun, Jean Ponce.

This work proposes a novel approach to learning mid-level image models for image categorization and cosegmentation. We represent each image class by a dictionary of discriminative part detectors that best discriminate that class from the background. We learn category-specific part detectors in a weakly supervised setting in which the training images are only labeled with category labels without part / object location labels. We use a latent SVM model regularized by l1,2 group sparsity to learn the discriminative 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 learned part detectors to image classification and cosegmentation, and quantitative experiments with standard benchmarks show that our approach matches or improves upon the state of the art. This work has been submitted to PAMI [24].

6.4. Image restoration, manipulation and enhancement

6.4.1. Fast Local Laplacian Filters: Theory and Applications

Participants: Mathieu Aubry, Sylvain Paris [Adobe], Samuel Hasinoff [Google], Jan Kautz [University College London], Fredo Durand [MIT].

Multi-scale manipulations are central to image editing but they are also prone to halos. Achieving artifact-free results requires sophisticated edge-aware techniques and careful parameter tuning. These shortcomings were recently addressed by the local Laplacian filters, which can achieve a broad range of effects using standard Laplacian pyramids. However, these filters are slow to evaluate and their relationship to other approaches is unclear. In this work, we show that they are closely related to anisotropic diffusion and to bilateral filtering. Our study also leads to a variant of the bilateral filter that produces cleaner edges while retaining its speed. Building upon this result, we describe an acceleration scheme for local Laplacian filters on gray-scale images that yields speed-ups on the order of 50x. Finally, we demonstrate how to use local Laplacian filters to alter the distribution of gradients in an image. We illustrate this property with a robust algorithm for photographic style transfer. This work has been published at ACM Transactions on Graphics 2014 [2].

6.4.2. Learning a Convolutional Neural Network for Non-uniform Motion Blur Removal

Participants: Jian Sun, Wenfei Cao, Zongben Xu, Jean Ponce.

In work work, we address the problem of estimating and removing non-uniform motion blur from a single blurry image. We propose a deep learning approach to predicting the probabilistic distribution of motion blur at the patch level using a convolutional neural network (CNN). We further extend the candidate set of motion kernels predicted by the CNN using carefully designed image rotations. A Markov random field model is then used to infer a dense non-uniform motion blur field enforcing the motion smoothness. Finally the motion blur is removed by a non-uniform deblurring model using patch-level image prior. Experimental evaluations show that our approach can effectively estimate and remove complex non-uniform motion blur that cannot be well achieved by the previous approaches. This work has been submitted to CVPR 2015.

6.5. Human activity capture and classification

6.5.1. Weakly Supervised Action Labeling in Videos Under Ordering Constraints

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

We are given a set of video clips, each one annotated with an ordered list of actions, such as "walk" then "sit" then "answer phone" extracted from, for example, the associated text script. We seek to temporally localize the individual actions in each clip as well as to learn a discriminative classifier for each action. We formulate the problem as a weakly supervised temporal assignment with ordering constraints. Each video clip is divided into small time intervals and each time interval of each video clip is assigned one action label, while respecting the order in which the action labels appear in the given annotations. We show that the action label assignment can be determined together with learning a classifier for each action in a discriminative manner. We evaluate the proposed model on a new and challenging dataset of 937 video clips with a total of 787720 frames containing sequences of 16 different actions from 69 Hollywood movies. This work has been published at ECCV 2014 [10].

6.5.2. Predicting Actions from Static Scenes

Participants: Tuan-Hung Vu, Catherine Olsson [MIT], Ivan Laptev, Aude Oliva [MIT], Josef Sivic.

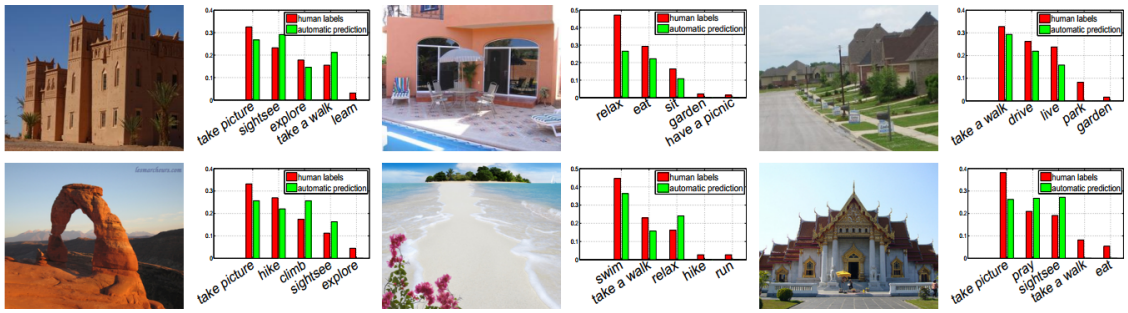

Figure 7. Automatic visual action prediction for test images in SUN Action dataset.

Human actions naturally co-occur with scenes. In this work we aim to discover action-scene correlation for a large number of scene categories and to use such correlation for action prediction. Towards this goal, we collect a new SUN Action dataset with manual annotations of typical human actions for 397 scenes. We next discover action-scene associations and demonstrate that scene categories can be well identified from their associated actions. Using discovered associations, we address a new task of predicting human actions for images of static scenes. We evaluate prediction of 23 and 38 action classes for images of indoor and outdoor scenes respectively and show promising results, see Figure 7. We also propose a new application of geo-localized action prediction and demonstrate ability of our method to automatically answer queries such as "Where is a good place for a picnic?" or "Can I cycle along this path?". This work has been published in ECCV 2014 [16].

6.5.3. Efficient feature extraction, encoding and classification for action recognition

Participants: Vadim Kantorov, Ivan Laptev.

Local video features provide state-of-the-art performance for action recognition. While the accuracy of action recognition has been continuously improved over the recent years, the low speed of feature extraction and subsequent recognition prevents current methods from scaling up to real-size problems. We address this issue and first develop highly efficient video features using motion information in video compression. We next explore feature encoding by Fisher vectors and demonstrate accurate action recognition using fast linear classifiers. Our method improves the speed of video feature extraction, feature encoding and action classification by two orders of magnitude at the cost of minor reduction in recognition accuracy. We validate our approach and compare it to the state of the art on four recent action recognition datasets. This work has been published at CVPR 2014 [12].

6.5.4. On Pairwise Cost for Multi-Object Network Flow Tracking

Participants: Visesh Chari, Simon Lacoste-Julien [Inria Sierra], Ivan Laptev, Josef Sivic.

Multi-object tracking has been recently approached with the min-cost network flow optimization techniques. Such methods simultaneously resolve multiple object tracks in a video and enable modeling of dependencies among tracks. Min-cost network flow methods also fit well within the "tracking-by-detection" paradigm where object trajectories are obtained by connecting per-frame outputs of an object detector. Object detectors, however, often fail due to occlusions and clutter in the video. To cope with such situations, we propose an approach that regularizes the tracker by adding second order costs to the min-cost network flow framework. While solving such a problem with integer variables is NP-hard, we present a convex relaxation with an efficient rounding heuristic which empirically gives certificates of small suboptimality. Results are shown on real world video sequences and demonstrate that the new constraints help selecting longer and more accurate tracks improving over the baseline tracking-by-detection method. This work has been submitted to CVPR 2015 [21].

7. Bilateral Contracts and Grants with Industry

7.1. 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 2018 video will account for 80-90% 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.

7.2. 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. 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.

8.2.2. 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.

8.2.3. European Research Council (ERC) Starting Grant: "Leap" - Josef Sivic

Participant: Josef Sivic.

The contract has begun on Nov 1st 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.

8.2.4. 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).

8.3. International Initiatives

8.3.1. IARPA FINDER Visual geo-localization (Inria)

Participants: Josef Sivic, Petr Gronat, Relja Arandjelovic.

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 / UC Berkeley). 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 during three months in 2014.

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 one month in 2014. Prof. John Canny (UC Berkeley) has visited Willow during three months in 2014 within the framework of Inria’s International Chair program.

8.4.1.1. Internships

Stefan Lee (Indiana University) has been a visiting PhD student at Willow since May 2014. Yumin Suh (Seoul National University) has been a visiting PhD student at Willow since Dec. 2014.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. General chair, scientific chair

- J. Sivic is a program co-chair of IEEE International Conference on Computer Vision (ICCV), 2015.

9.1.1.2. Member of the organizing committee

- Workshop co-organizer, THUMOS Challenge 2014: Action Recognition with a Large Number of Classes, in conjunction with RCCV'14, Zurich, Switzerland (Ivan Laptev).

9.1.2. Scientific events selection

9.1.2.1. Area chairs

- European Conference on Computer Vision (ECCV), 2014 (I. Laptev).
- Asian Conference on Computer Vision (ACCV), 2014 (I. Laptev).
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015 (I. Laptev, J. Ponce).

9.1.2.2. Member of the conference program committee

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014 (I. Laptev, M. Cho, J. Sivic).
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015 (R. Arandjelović, M. Cho, J. Sivic).
- European Conference on Computer Vision (ECCV), 2014 (R. Arandjelović, M. Cho, J. Sivic).
- Asian Conference on Computer Vision (ACCV), 2014 (M. Cho, S. Kwak, J. Sun).
- 29th Conference on Artificial Intelligence (AAAI), 2015 (I. Laptev).
- 5th Int Workshop on Human Behavior Understanding (HBU), 2014 (I. Laptev).

9.1.3. Journal

9.1.3.1. Member of the editorial board

- International Journal of Computer Vision (I. Laptev, J. Ponce, J. Sivic).
- IEEE Transactions on Pattern Analysis and Machine Intelligence (I. Laptev).
- Foundations and Trends in Computer Graphics and Vision (J. Ponce).
- Image and Vision Computing Journal (I. Laptev).
- 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.

9.1.3.2. Reviewer

- International Journal of Computer Vision (R. Arandjelović, P. Bojanowski, M. Cho, S. Kwak, J. Sun).
- IEEE Transactions on Pattern Analysis and Machine Intelligence (R. Arandjelović, M. Cho, S. Kwak).
- IEEE Transactions on Circuits and Systems for Video Technology (B. Ham, S. Kwak).
- IEEE Transactions on Image Processing (B. Ham, J. Sun).
- IEEE Signal Processing Letters (B. Ham).
- Computer Vision and Image Understanding (M. Cho).

- Pattern Recognition (R. Arandjelović).
- Pattern Recognition Letters (B. Ham).
- Signal Processing: Image Communication (B. Ham).
- EURASIP Journal on Image and Video Processing (B. Ham).

9.2. Teaching - Supervision - Juries

9.2.1. HdR

HdR: Josef Sivic, "Visual recognition of objects, people and places", École normale supérieure Paris.

9.2.2. Teaching

Licence : J. Ponce, "Introduction to computer vision", 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.

Master : I. Laptev, J. Ponce and J. Sivic (together with Z. Harchaoui and J. Mairal, Inria Grenoble and F. Bach), Cours PSL-ITI - Informatique, mathématiques appliquées pour le traitement du signal et l'imagerie, 20h.

Master: I. Laptev, course "Visual object recognition and localization", SkolTech, Moscow, January 2014, 30h.

Doctorat: I. Laptev, Tutorial on human action recognition, CVPR'14 tutorial "Emerging topics in human activity recognition".

Doctorat: I. Laptev, course "Visual Recognition: Objects, Actions and Scenes", University of Trento, Italy, July 2014, 20h.

9.2.3. Supervision

PhD in progress : Jean-Baptiste Alayrac, "Structured learning from video and natural language", started in 2014, I. Laptev, J. Sivic and S. Lacoste-Julien.

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 : 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 : Guilhem Chŕon, "Structured modeling and recognition of human actions in video", started in 2014, I. Laptev and C. Schmid.

PhD in progress : Théophile Dalens, "Learning to analyze and reconstruct architectural scenes", starting 1 Jan 2015, M. Aubry and J. Sivic.

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, I. Laptev and J. Sivic.

PhD in progress : Matthew Trager, “Projective geometric models in vision”, started in 2014, J. Ponce and M. Hebert (CMU).

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

9.2.4. Juries

- PhD thesis committee:
  - Antoine Fagette, UPMC, 2014 (I. Laptev, rapporteur).
  - Ivo Everts, Amsterdam University, 2014 (I. Laptev).
  - Piotr Bilinski, University of Nice, 2014 (I. Laptev).
  - Laurent Sifre, Ecole Polytechnique, 2014 (J. Sivic).
  - Omid Aghazadeh, KTH Stockholm, 2014 (J. Sivic).
- HDR thesis committee:
  - Josef Sivic, ENS Ulm, 2014 (J. Ponce).
  - Herve Jegou, Université de Rennes, 2014 (J. Ponce).
- Other:
  - Member of the PSL Research Council, 2012- (J. Ponce).
  - Member of the Mathematics and Systems Unit AERES evaluation committee at Mines ParisTech, January 2014 (J. Ponce).
  - Member of the faculty selection committee, Université de Marne la Vallée, 2014 (J. Ponce).
  - Head, AERES committee in charge of CMLA’s evaluation at ENS Cachan, 2014 (J. Ponce).
  - Member of the ESF/FCT evaluation committee for Portuguese research centers, 2014 (J. Ponce).
  - Member of the AERES evaluation committee at Ecole Centrale de Paris, 2014 (I. Laptev).
  - Member of the Inria postdoc selection committee, 2012- (I. Laptev).
  - Member of Inria Commission de developpement technologique (CDT), 2012- (J. Sivic).

9.3. Invited presentations

- M. Cho, Invited talk, Ecole des Ponts ParisTech, Paris, France, Jan. 2014.
- I. Laptev, Plenary speaker, ICVGIP, Dec. 14-17, Bangalore, India, 2014.
- I. Laptev, Invited speaker, Inria Sophia-Antipolis, Dec. 5, 2014.
- I. Laptev, Invited speaker, KU Leuven, Belgium, November 24, 2014.
- I. Laptev, Invited speaker, Steklov Institute of Mathematics, Saint Petersburg, Russia, Nov. 17, 2014.
- I. Laptev, Invited speaker, Aalto Univ., Helsinki, Finland, Aug. 21, 2014.
- I. Laptev, Invited speaker, ECCV’14 Area Chair Workshop, Zurich, Switzerland, June 2014.
- I. Laptev, Invited speaker, CVPR’14 Workshop on Perceptual Organization, Columbus, USA, June 2014.
- I. Laptev, Invited speaker, Idiap Research Institute, Martigny, Switzerland, Feb. 2014.
- I. Laptev, Invited speaker, EPFL, Lausanne, Switzerland, Feb. 2014.
- I. Laptev, Invited speaker, Computer Vision Winter Workshop, Křtiny, Czech Republic, Feb. 2014.
- I. Laptev, Invited speaker, SkolTech, Moscow, Russia, Jan. 2014.
- I. Laptev, Invited speaker, Royal Institute of Technology, Stockholm, Sweden, Jan. 2014.

- J. Ponce, Invited speaker, Nanyang Technological University, Singapore, Dec. 2014.
- J. Ponce, Keynote speaker, 3DV conference, Tokyo, Dec. 2014.
- J. Ponce, Plenary speaker, European Workshop on Visual Information Processing, Paris, Dec. 2014.
- J. Sivic, Invited talk at the Center for Research in Computer Vision, University of Central Florida, April 2014.
- J. Sivic, Seminar, New York University, April 2014.
- J. Sivic, Seminar, University of Oxford, April 2014.
- J. Sivic, Seminar, KTH Stockholm, June 2014.
- J. Sivic, Seminar, Inria Sophia-Antipolis, July 2014.
- J. Sivic, Seminar, MPI Saarbruecken, July 2014.
- J. Sivic, Seminar, UC Berkeley, August 2014.
- J. Sivic, Seminar, Adobe Research, August 2014.
- J. Sivic, Talk at an invited workshop, CIFAR, Montreal, December 2014.
- J. Sivic, Invited talk, Smart City workshop, Paris, June 2014.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] J. SIVIC. *Visual search and recognition of objects, scenes and people*, Ecole Normale Supérieure de Paris - ENS Paris, February 2014, Habilitation à diriger des recherches, https://tel.archives-ouvertes.fr/tel-01064559

Articles in International Peer-Reviewed Journals

[2] M. AUBRY, S. PARIS, S. HASINOFF, J. KAUTZ, F. DURAND. *Fast Local Laplacian Filters: Theory and Applications*, in "ACM Transactions on Graphics", 2014, vol. 33, n^{o} 5, pp. 167.1-167.14 [*DOI :* 10.1145/2629645], https://hal.archives-ouvertes.fr/hal-01063419

[3] M. AUBRY, B. C. RUSSELL, J. SIVIC. *Painting-to-3D Model Alignment Via Discriminative Visual Elements*, in "ACM Transactions on Graphics", March 2014, vol. 33, n^{o} 2 [*DOI :* 10.1145/2591009], https://hal.inria.fr/hal-00863615

[4] D. FOUHEY, V. DELAITRE, A. GUPTA, A. EFROS, I. LAPTEV, J. SIVIC. *People Watching: Human Actions as a Cue for Single View Geometry*, in "International Journal of Computer Vision", February 2014, https://hal.inria.fr/hal-01066257

[5] B. GOLDLUECKE, M. AUBRY, K. KOLEV, D. CREMERS. *A Super-resolution Framework for High-Accuracy Multiview Reconstruction*, in "International Journal of Computer Vision", January 2014, vol. 106, n^{o} 2, pp. 172-191 [*DOI :* 10.1007/s11263-013-0654-8], https://hal.inria.fr/hal-01057502

[6] G. SEGUIN, K. ALAHARI, J. SIVIC, I. LAPTEV. *Pose Estimation and Segmentation of Multiple People in Stereoscopic Movies*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2014, 1 p. [*DOI :* 10.1109/TPAMI.2014.2369050], https://hal.inria.fr/hal-01089660

[7] O. WHYTE, J. SIVIC, A. ZISSERMAN. *Deblurring Shaken and Partially Saturated Images*, in "International Journal of Computer Vision", 2014 [*DOI :* 10.1007/s11263-014-0727-3], https://hal.inria.fr/hal-01053888

International Conferences with Proceedings

[8] M. ANDREUX, E. RODOLA, M. AUBRY, D. CREMERS. *Anisotropic Laplace-Beltrami Operators for Shape Analysis*, in "NORDIA'14 - Sixth Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment", Zurich, Switzerland, September 2014, https://hal.inria.fr/hal-01057244

[9] M. AUBRY, D. MATURANA, A. EFROS, B. RUSSELL, J. SIVIC. *Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models*, in "CVPR 2014 - Compter Vision and Pattern Recognition", Colombus OH, United States, June 2014, https://hal.inria.fr/hal-01057240

[10] P. BOJANOWSKI, R. LAJUGIE, F. BACH, I. LAPTEV, J. PONCE, C. SCHMID, J. SIVIC. *Weakly Supervised Action Labeling in Videos Under Ordering Constraints*, in "ECCV - European Conference on Computer Vision", Zurich, Switzerland, September 2014, pp. 628-643 [*DOI :* 10.1007/978-3-319-10602-1_41], https://hal.inria.fr/hal-01053967

[11] M. CHO, J. SUN, O. DUCHENNE, J. PONCE. *Finding Matches in a Haystack: A Max-Pooling Strategy for Graph Matching in the Presence of Outliers*, in "CVPR - IEEE Conference on Computer Vision and Pattern Recognition", Columbus, Ohio, United States, June 2014, https://hal.inria.fr/hal-01053675

[12] V. KANTOROV, I. LAPTEV. *Efficient feature extraction, encoding and classification for action recognition*, in "CVPR 2014 - Computer Vision and Pattern Recognition", Columbus, United States, June 2014, https://hal.inria.fr/hal-01058734

[13] M. OQUAB, L. BOTTOU, I. LAPTEV, J. SIVIC. *Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks*, in "IEEE Conference on Computer Vision and Pattern Recognition", Columbus, OH, United States, June 2014, Conference version of the paper, https://hal.inria.fr/hal-00911179

[14] J. PONCE, M. HEBERT. *On Image Contours of Projective Shapes*, in "ECCV - European Conference on Computer Vision", Zurich, Switzerland, September 2014, https://hal.inria.fr/hal-01053677

[15] J. PONCE, M. HÉBERT. *Trinocular Geometry Revisited*, in "CVPR - IEEE Conference on Computer Vision and Pattern Recognition", Columbus, Ohio, United States, June 2014, https://hal.inria.fr/hal-01053676

[16] T.-H. VU, C. OLSSON, I. LAPTEV, A. OLIVA, J. SIVIC. *Predicting Actions from Static Scenes*, in "ECCV'14 - 13th European Conference on Computer Vision", Zurich, Switzerland, D. FLEET, T. PAJDLA, B. SCHIELE, T. TUYTELAARS (editors), Springer, September 2014, vol. 8693, pp. 421-436 [*DOI :* 10.1007/978-3-319-10602-1_28], https://hal.inria.fr/hal-01053935

National Conferences with Proceedings

[17] M. AUBRY, B. C. RUSSELL, J. SIVIC. *Where was this picture painted ? - Localizing paintings by alignment to 3D models*, in "Reconnaissance de Formes et Intelligence Artificielle (RFIA) 2014", Rouen, France, June 2014, 6 p. , https://hal.archives-ouvertes.fr/hal-00988830

Scientific Books (or Scientific Book chapters)

[18] J. MAIRAL, F. BACH, J. PONCE. *Sparse Modeling for Image and Vision Processing*, Foundations and Trends in Computer Graphics and Vision, now publishers, December 2014, vol. 8, n^{o} 2-3, 216 p. [*DOI :* 10.1561/9781680830095], https://hal.inria.fr/hal-01081139

[19] O. WHYTE, J. SIVIC, A. ZISSERMAN, J. PONCE. *Efficient, Blind, Spatially-Variant Deblurring for Shaken Images*, in "Motion Deblurring: Algorithms and Systems", A. N. RAJAGOPALAN, R. CHELLAPPA (editors), Cambridge University Press, 2014, https://hal.inria.fr/hal-01063814

Research Reports

[20] T. DALENS, J. SIVIC, I. LAPTEV, M. CAMPEDEL. *Painting recognition from wearable cameras*, September 2014, 13 p. , https://hal.inria.fr/hal-01062126

Other Publications

[21] V. CHARI, S. LACOSTE-JULIEN, I. LAPTEV, J. SIVIC. *On Pairwise Cost for Multi-Object Network Flow Tracking*, August 2014, https://hal.inria.fr/hal-01110678

[22] M. CHO, S. KWAK, C. SCHMID, J. PONCE. *Unsupervised Object Discovery and Localization in the Wild: Part-based Matching with Bottom-up Region Proposals*, January 2015, https://hal.inria.fr/hal-01110036

[23] M. OQUAB, L. BOTTOU, I. LAPTEV, J. SIVIC. *Weakly supervised object recognition with convolutional neural networks*, June 2014, https://hal.inria.fr/hal-01015140

[24] J. SUN, J. PONCE. *Learning Dictionary of Discriminative Part Detectors for Image Categorization and Cosegmentation*, September 2014, https://hal.archives-ouvertes.fr/hal-01064637