IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2015

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

Creation of the Project-Team: 2007 June 01

Keywords:

Computer Science and Digital Science:

3.1.1. - Modeling, representation
3.4. - Machine learning and statistics
5.3. - Image processing and analysis
5.4. - Computer vision
8. - Artificial intelligence
8.1. - Knowledge
8.2. - Machine learning

Other Research Topics and Application Domains:

9.4.1. - Computer science
9.4.5. - Data science

1. Members

Research Scientists

John Canny [Inria, International Chair, UC Berkeley]
Minsu Cho [Inria, Starting Research position]
Ivan Laptev [Inria, Senior Researcher, HdR]
Josef Sivic [Inria, Researcher, HdR]

Faculty Member

Jean Ponce [Team leader, ENS Paris, Professor]

Engineers

Andrei Bursuc [Inria]
Jonathan Chemla [Inria, from Mar 2015]
Vincent Delaitre [Inria, until Nov 2015]
Petr Gronat [Inria, granted by ObjectVideo Inc]
Antony Marion [Inria, from Nov 2015]
Anastasia Syromyatnikova [Inria, until May 2015, granted by ERC (European Research Council Excecutive Agency)]

PhD Students

Piotr Bojanowski [Inria, granted by FP7 VideoWorld project]
Guilhem Cheron [Inria]
Florent Couzinie-Devy [ENS Cachan, until Mar 2015]
Théophile Dalens [Inria]
Vadim Kantorov [Inria, granted by ERC (European Research Council Excecutive Agency)]
Maxime Oquab [Inria]
Julia Peyre [Inria, from Apr 2015]
Rafael Sampaio de Rezende [Inria, granted by FP7 VideoWorld project]
Guillaume Seguin [Inria]
Matthew Trager [Inria]
Gül Varol [Inria, from May 2015]
Tuan Hung Vu [Inria, granted by ERC (European Research Council Excecutive Agency)]

Post-Doctoral Fellows

Relja Arandjelovic [Inria]
Bumsub Ham [Inria]
Suha Kwak [Inria]

Visiting Scientists

Alyosha Efros [University of Berkeley, until Jul 2015]
Yasutaka Furukawa [Washington University, Jul 2015]
Vincent Lepetit [EPFL, Aug 2015]

Administrative Assistant

David Dinis [Inria]

Others

Yumin Suh [Seoul National Univ., until Jun 2015]
Nishant Agrawal [IIIT, India, from May 2015 until July 2015]
Mathieu Aubry [ENPC]
Bastien Jacquet [Swiss Federal Institute of Technology Zürich (ETH), until Aug 2015]
Filip Srajer [Czech Technical University, Feb 2015]
Michail Nikita [Moscow State Univ., Russia, Oct 2015]

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: Théophile Dalens (Inria), Julia Peyre (Inria), and Gül Varol (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, https://github.com/pmoulon/CMVS-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 7.1, are focused on: (i) continuing our theoretical study of multi-view camera geometry [17], [25]. (ii) modelling new representations of large-scale visual place recognition in structured image collections of urban environments [16], and (iii) developing new weakly supervised and deep learning approaches to large scale place recognition and retrieval [21].

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 7.2, has focused on: (i) learning object representation in a weakly supervised manner using convolutional neural networks [14], (ii) localizing objects and their parts from images and videos with minimum supervision [8], [11], (iii) discovering and analyzing architectural style elements from huge collections of street-level imagery [13], and (iv) developing new approaches to visual correspondence and scene flow using multi-scale region proposals and features [22].

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

3.3. 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 7.3, has focused on: (i) developing new image filters using both static and dynamic guidances in a unified optimization framework [10], and (ii) developing new formulations for image deblurring and restoration cast as a deep learning problem [15].

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 7.4.

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 have recently explored automatic mining of action categories and actor names from videos and corresponding scripts [6]. Within the PhD of Piotr Bojanowski and Jean-Baptiste Alayrac we extend this direction by modeling the temporal order of actions and developing models for learning key steps from instruction videos [20].

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. In particular, we develop deep learning methods and design new trainable representations for human action recognition. Such methods range from generic video-level representations based on space-time convolutional neural networks [27] to person-focused representations based on human pose [9], [23]. We also address the tasks of person detection [18], segmentation [4], [26] and tracking [7] in challenging video data.

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. Highlights of the Year

5.1. Highlights of the Year

J. Sivic has served as a Program Chair for International Conference on Computer Vision, Santiago, Chile, 2015

6. New Software and Platforms

6.1. Visual Place Recognition with Repetitive Structures

A new version of the open-source release of the software package for visual localization in urban environments has been made publicly available in July 2015. The software package implements the method [A. Torii et al., CVPR 2013] (journal version published this year in [5]) 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 first version was made available in 2013 and has been updated in May 2014. The current version of the software is available at http://www.di.ens.fr/willow/research/repttile/download/repttile_demo_ver04.zip.

6.2. NetVLAD: CNN architecture for weakly supervised place recognition

Open source release of the software package for our paper "NetVLAD: CNN architecture for weakly supervised place recognition" [21]. It provides a full implementation of the method, including code for weakly supervised training of the CNN representation, testing on standard datasets, as well as trained models. Links to all of these are available at our project page http://www.di.ens.fr/willow/research/netvlad/.

6.3. 24/7 place recognition by view synthesis

Open source release of the software package for our paper "24/7 place recognition by view synthesis" [16]. It provides code for computing VLAD descriptors, performing feature matching and view synthesis. Link to the code is available at our project page http://www.ok.ctrl.titech.ac.jp/~torii/project/247/.

6.4. Weakly Supervised Object Recognition with Convolutional Neural Networks

Open-source release of the software package for weakly supervised object recognition with convolutional neural networks has been made publicly available in May 2015. The software package implements the method [M. Oquab et al., CVPR 2015] [14] for object category recognition and localization using convolutional neural networks with weak supervision (without bounding box annotations). The method (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. The current version of the software is available at http://www.di.ens.fr/willow/research/weakcnn/.

6.5. Unsupervised Object Discovery and Localization in the Wild

This package contains source code for unsupervised object discovery and localizatoin from image collections. From an arbitrary collection of images in the wild, the method effectively discover dominant object instances and localize them by bounding boxes. The localization accuracy of discovered objects measured at standard benchmarks for object localization is significantly better than the state-of-the-art methods in co-localization, while using no supervision on image collections. The package is available from http://www.di.ens.fr/willow/research/objectdiscovery/.

6.6. Joint Static and Dynamic Guidance Filter

Open-source release of the software package for depth upsampling, texture removal, and scale-space filtering has been made publicly available. The software package implements the newly developed method [10] for robust filtering with joint static and dynamic guidance. The software is available at http://www.di.ens.fr/willow/research/sdfilter/.

7. New Results

7.1. 3D object and scene modeling, analysis, and retrieval

7.1.1. The joint image handbook

Participants: Matthew Trager, Martial Hebert, Jean Ponce.

Given multiple perspective photographs, point correspondences form the “joint image”, effectively a replica of three-dimensional space distributed across its two-dimensional projections. This set can be characterized by multilinear equations over image coordinates, such as epipolar and trifocal constraints. In this work, we revisit the geometric and algebraic properties of the joint image, and address fundamental questions such as how many and which multilinearities are necessary and/or sufficient to determine camera geometry and/or image correspondences. Our new theoretical results answer these questions in a very general setting, and our work, published ICCV 2015 [17], is intended to serve as a “handbook” reference about multilinearities for practitioners.

7.1.2. *Trinocular Geometry Revisited*

Participants: Jean Ponce, Martial Hebert, Matthew Trager.

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, which can be used to construct a practical and efficient method for trinocular geometry parameter estimation. This work has been published at CVPR 2014, and a revised version that includes numerical experiments using synthetic and real data has been submitted to IJCV [25].

7.1.3. *24/7 place recognition by view synthesis*

Participants: Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, Tomas Pajdla.

We address the problem of large-scale visual place recognition for situations where the scene undergoes a major change in appearance, for example, due to illumination (day/night), change of seasons, aging, or structural modifications over time such as buildings built or destroyed. Such situations represent a major challenge for current large-scale place recognition methods. This work has the following three principal contributions. First, we demonstrate that matching across large changes in the scene appearance becomes much easier when both the query image and the database image depict the scene from approximately the same viewpoint. Second, based on this observation, we develop a new place recognition approach that combines (i) an efficient synthesis of novel views with (ii) a compact indexable image representation. Third, we introduce a new challenging dataset of 1,125 camera-phone query images of Tokyo that contain major changes in illumination (day, sunset, night) as well as structural changes in the scene. We demonstrate that the proposed approach significantly outperforms other large-scale place recognition techniques on this challenging data. This work has been published at CVPR 2015 [16]. Figure 1 shows examples of the newly collected Tokyo 24/7 dataset.

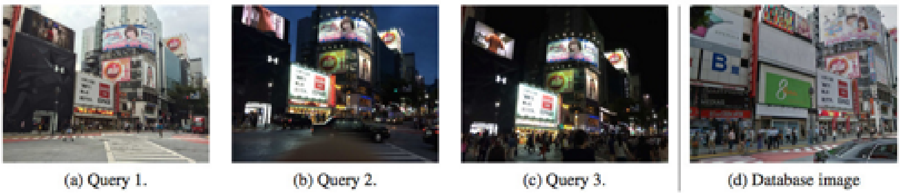
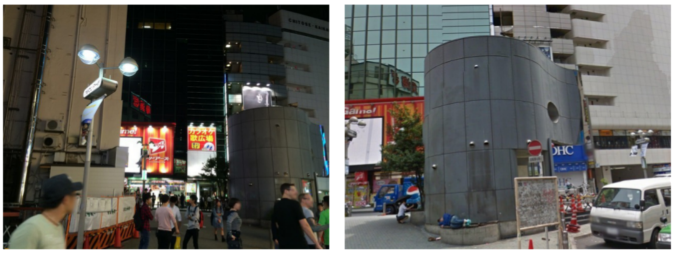

(a) Query 1. (b) Query 2. (c) Query 3. (d) Database image

Figure 1. Example query images from the newly collected 24/7 Tokyo dataset. Each place in the query set is captured at different times of day: (a) daytime, (b) sunset, and (c) night. For comparison, the database street-view image at a close-by position is shown in (d). Note the major changes in appearance (illumination changes in the scene) between the database image (d) and the query images (a,b,c)

7.1.4. *NetVLAD: CNN architecture for weakly supervised place recognition*

Participants: Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, Josef Sivic.

In [21], we tackle the problem of large scale visual place recognition, where the task is to quickly and accurately recognize the location of a given query photograph. We present the following three principal contributions. First, we develop a convolutional neural network (CNN) architecture that is trainable in an end-to-end manner directly for the place recognition task. The main component of this architecture, NetVLAD, is a new generalized VLAD layer, inspired by the "Vector of Locally Aggregated Descriptors" image representation commonly used in image retrieval. The layer is readily pluggable into any CNN architecture and amenable to training via backpropagation. Second, we develop a training procedure, based on a new weakly supervised ranking loss, to learn parameters of the architecture in an end-to-end manner from images depicting the same places over time downloaded from Google Street View Time Machine. Finally, we show that the proposed architecture obtains a large improvement in performance over non-learnt image representations as well as significantly outperforms off-the-shelf CNN descriptors on two challenging place recognition benchmarks. This work is under review. Figure 2 shows some qualitative results.



(a) Mobile phone query (b) Retrieved image of same place

Figure 2. Our trained NetVLAD descriptor correctly recognizes the location (b) of the query photograph (a) despite the large amount of clutter (people, cars), changes in viewpoint and completely different illumination (night vs daytime).

7.2. Category-level object and scene recognition

7.2.1. *Is object localization for free? – Weakly-supervised learning with convolutional neural networks*

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

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 3). 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 published at CVPR 2015 [14] . Illustration of localization results by our method in Microsoft COCO dataset is shown in Figure 4.

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

Participants: Minsu Cho, Suha Kwak, Cordelia Schmid, Jean Ponce.

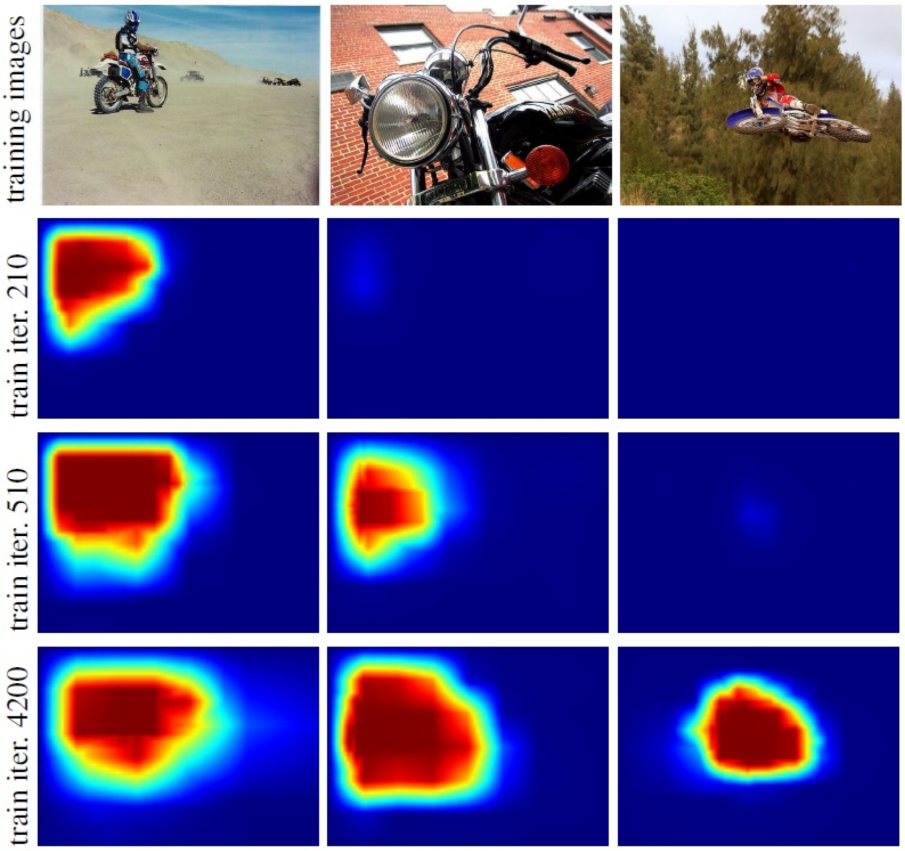


Figure 3. Evolution of localization score maps for the motorbike class over iterations of our weakly-supervised CNN training. Note that locations of objects with more usual appearance are discovered earlier during training.

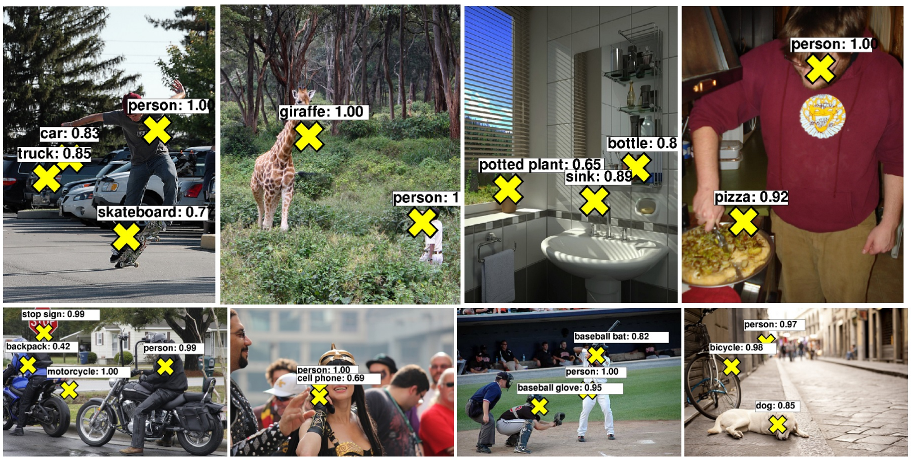


Figure 4. Example location predictions for images from the Microsoft COCO validation set obtained by our weakly-supervised method. Note that our method does not use object locations at training time, yet can predict locations of objects in test images (yellow crosses). The method outputs the most confident location for most confident object classes.

In [8], we address *unsupervised* discovery and localization of dominant objects from a noisy image collection of multiple object classes. The setting of this problem is fully unsupervised (Fig. 5), 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 unsupervised discovery and localization problem using a part-based region 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 for each candidate correspondence 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 evaluations on standard benchmarks (e.g., Object Discovery and PASCAL VOC 2007 datasets) demonstrate that the proposed approach significantly outperforms the current state of the art in colocalization, and achieves robust object discovery even in a fully unsupervised setting. This work has been published in CVPR 2015 [8] as oral presentation.

7.2.3. Unsupervised Object Discovery and Tracking in Video Collections

Participants: Suha Kwak, Minsu Cho, Ivan Laptev, Jean Ponce, Cordelia Schmid.

In [11], we address the problem of automatically localizing dominant objects as spatio-temporal tubes in a noisy collection of videos with minimal or even no supervision. We formulate the problem as a combination of two complementary processes: discovery and tracking (Figure 6). The first one establishes correspondences bet ween prominent regions across videos, and the second one associates similar object regions within the same video. It is empirically demonstrated that our method can handle video collections featuring multiple object classes, and substantially outperforms the state of the art in colocalization, even though it tackles a broader problem with much less supervision. This work has been published in ICCV 2015.

7.2.4. Linking Past to Present: Discovering Style in Two Centuries of Architecture

Participants: Stefan Lee, Nicolas Maisonneuve, David Crandall, Alexei A. Efros, Josef Sivic.

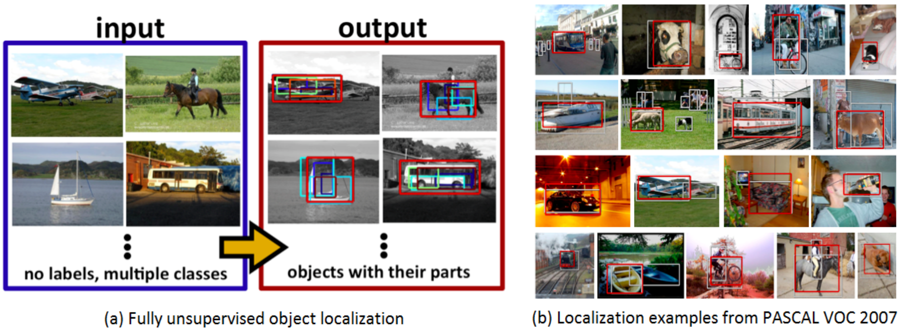

input
no labels, multiple classes
output
objects with their parts
(a) Fully unsupervised object localization
(b) Localization examples from PASCAL VOC 2007

Figure 5. Unsupervised object discovery in the wild. We tackle object localization in an unsupervised scenario without any type 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).

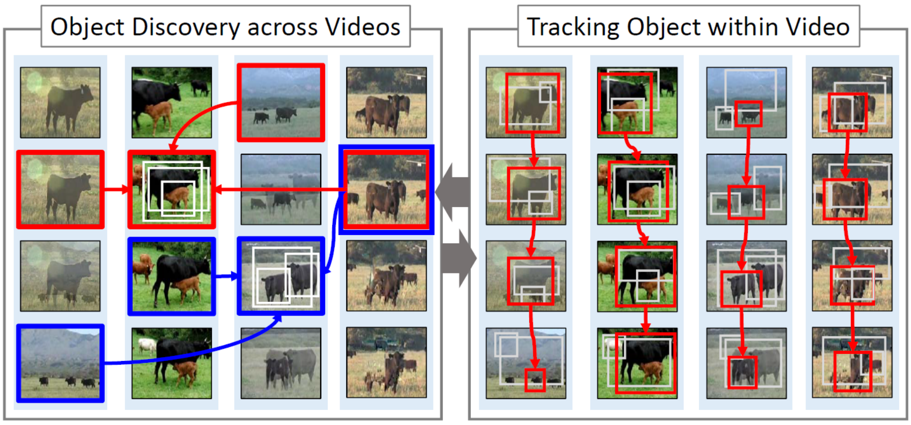


Figure 6. Dominant objects in a video collection are discovered by analyzing correspondences between prominent regions across videos (left). Within each video, object candidates, discovered by the former process, are temporally associated and a smooth spatio-temporal localization is estimated (right). These processes are alternated until convergence or up to a fixed number of iterations.

With vast quantities of imagery now available online, researchers have begun to explore whether visual patterns can be discovered automatically. Here we consider the particular domain of architecture, using huge collections of street-level imagery to find visual patterns that correspond to semantic-level architectural elements distinctive to particular time periods. We use this analysis both to date buildings, as well as to discover how functionally similar architectural elements (e.g. windows, doors, balconies, etc.) have changed over time due to evolving styles. We validate the methods by combining a large dataset of nearly 150,000 Google Street View images from Paris with a cadastre map to infer approximate construction date for each facade. Not only could our analysis be used for dating or geo-localizing buildings based on architectural features, but it also could give architects and historians new tools for confirming known theories or even discovering new ones. The work was published in [13] and the results are illustrated in figure 7.

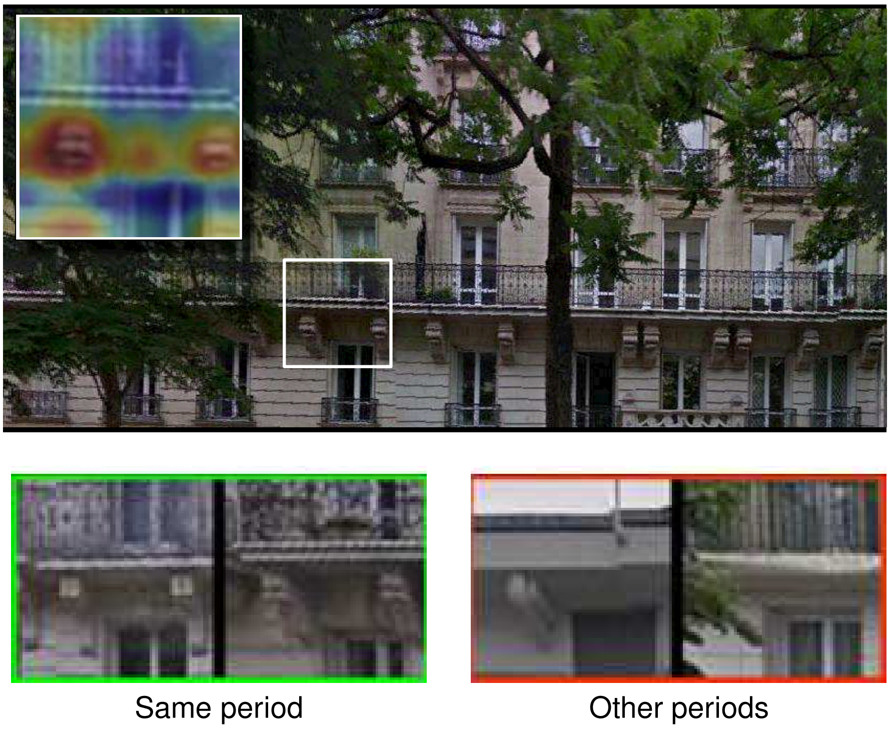


Figure 7. Using thousands of Street View images aligned to a cadastral map, we automatically find visual elements distinctive to particular architectural periods. For example, the patch in white above was found to be distinctive to the Haussmann period (late 1800's) in Paris, while the heat map (inset) reveals that the ornate balcony supports are the most distinctive features. We can also find functionally-similar elements fromthe same and different time periods (bottom).

7.2.5. Proposal Flow

Participants: Bumsub Ham, Minsu Cho, Cordelia Schmid, Jean Ponce.

Finding image correspondences remains a challenging problem in the presence of intra-class variations and large changes in scene layout, typical in scene flow computation. In [22], we introduce a novel approach to this problem, dubbed proposal flow, that establishes reliable correspondences using object proposals. Unlike prevailing scene flow approaches that operate on pixels or regularly sampled local regions, proposal flow benefits from the characteristics of modern object proposals, that exhibit high repeatability at multiple scales, and can take advantage of both local and geometric consistency constraints among proposals. We also show that proposal flow can effectively be transformed into a conventional dense flow field. We introduce a new dataset that can be used to evaluate both general scene flow techniques and region-based approaches such as proposal flow. We use this benchmark to compare different matching algorithms, object proposals, and region features within proposal flow with the state of the art in scene flow. This comparison, along with experiments on standard datasets, demonstrates that proposal flow significantly outperforms existing scene flow methods in various settings. This work is under review. The proposed method and its qualitative result are illustrated in Figure 8.

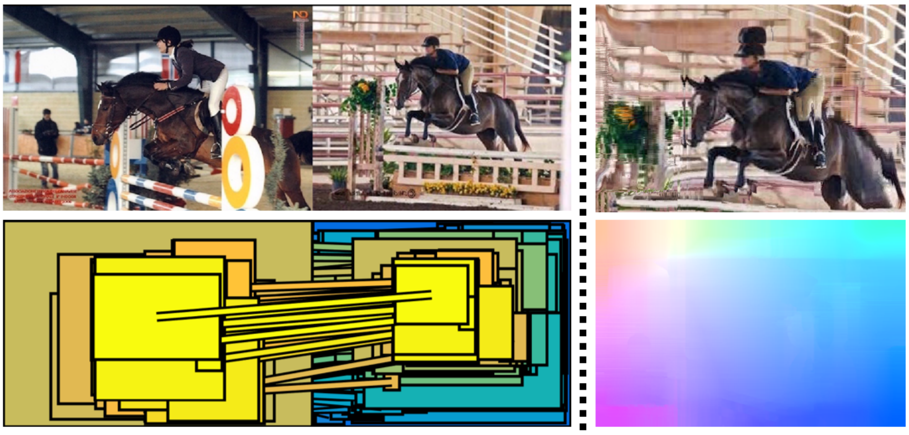


Figure 8. Proposal flow generates a reliable scene flow between similar images by establishing geometrically consistent correspondences between object proposals. (Left) Region-based scene flow by matching object proposals. (Right) Color-coded dense flow field generated from the region matches, and image warping using the flow.

7.3. Image restoration, manipulation and enhancement

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

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

In this 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 published at CVPR 2015[15].

7.3.2. Robust Image Filtering Using Joint Static and Dynamic Guidance

Participants: Bumsub Ham, Minsu Cho, Jean Ponce.

Filtering images using a guidance signal, a process called joint or guided image filtering, has been used in various tasks in computer vision and computational photography, particularly for noise reduction and joint upsampling. The aim is to transfer the structure of the guidance signal to an input image, restoring noisy or altered image structure. The main drawbacks of such a data-dependent framework are that it does not consider differences in structure between guidance and input images, and it is not robust to outliers. We propose a novel SD (for static/dynamic) filter to address these problems in a unified framework by jointly leveraging structural information of guidance and input images. Joint image filtering is formulated as a nonconvex optimization problem, which is solved by the majorization-minimization algorithm. The proposed algorithm converges quickly while guaranteeing a local minimum. The SD filter effectively controls the underlying image structure at different scales and can handle a variety of types of data from different sensors. It is robust to outliers and other artifacts such as gradient reversal and global intensity shifting, and has good edge-preserving smoothing properties. We demonstrate the flexibility and effectiveness of the SD filter in a great variety of applications including depth upsampling, scale-space filtering, texture removal, flash/non-flash denoising, and RGB/NIR denoising. This has been published at CVPR 2015 [10]. The SD filter is illustrated in Figure 9.

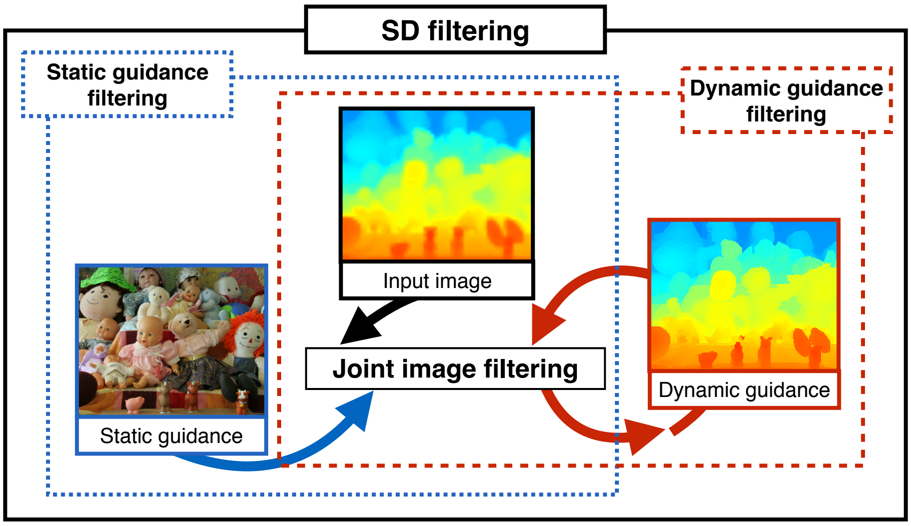


Figure 9. Sketch of joint image filtering and SD filtering: Static guidance filtering convolves an input image with a weight function computed from static guidance, as in the dotted blue box. Dynamic guidance filtering uses weight functions that are repeatedly obtained from regularized input images, as in the dotted red box. We have observed that static and dynamic guidance complement each other, and exploiting only one of them is problematic, especially in the case of data from different sensors (e.g., depth and color images). The SD filter takes advantage of both, and addresses the problems of current joint image filtering.

7.3.3. PCS-Net: A Deep learning approach to image restoration

Participants: Jian Sun, Jean Ponce.

This work introduces a novel framework for image restoration casting this problem as a joint classification and regression task. This is a learning-based approach, which first classifies degraded image patches into different categories, then restores these patches using category-specific models. We implement this idea by designing a novel convolutional neural network (dubbed PCS-Net), combining a CNN-based patch classification subnet with a novel patch category switched CNN architecture for category-specific restoration. The proposed PCS-Net learns different weights for different patch categories in a common network structure. Experiments on standard benchmarks show that our approach matches or improves upon the state of the art in image super-resolution and denoising. This work is under review.

7.4. Human activity capture and classification

7.4.1. P-CNN: Pose-based CNN Features for Action Recognition

Participants: Guilhem Chéron, Ivan Laptev, Cordelia Schmid.

This work [9] targets human action recognition in video. We argue for the importance of a representation derived from human pose. To this end we propose a new Pose-based Convolutional Neural Network descriptor (P-CNN) for action recognition. The descriptor aggregates motion and appearance information along tracks of human body parts as shown in Figure 10. We experiment with P-CNN features obtained both for automatically estimated and manually annotated human poses. We evaluate our method on JHMDB and MPII Cooking datasets. For both datasets our method shows consistent improvement over the state of the art. This work has been published at ICCV 2015 [9], and P-CNN code (Matlab) is available online at http://www.di.ens.fr/willow/research/p-cnn/.

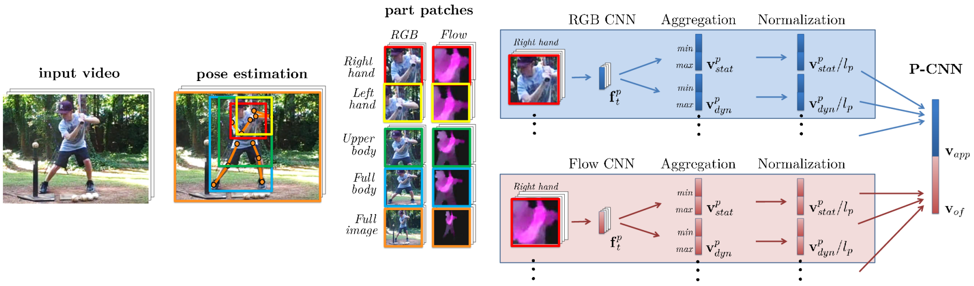


Figure 10. P-CNN features. From left to right: Input video. Human pose. Patches of appearance and optical flow for human body parts. One RGB and one flow CNN descriptor is extracted per frame and per part. Frame descriptors are aggregated over time to obtain the video descriptor. Video descriptors are normalized and concatenated into appearance features and flow features. The final P-CNN feature is the concatenation of appearance and flow.

7.4.2. Context-aware CNNs for person head detection

Participants: Tuan-Hung Vu, Anton Osokin, Ivan Laptev.

Person detection is a key problem for many computer vision tasks. While face detection has reached maturity, detecting people under a full variation of camera view-points, human poses, lighting conditions and occlusions is still a difficult challenge. In this work we focus on detecting human heads in natural scenes. Starting from the recent local R-CNN object detector, we extend it with two types of contextual cues. First, we leverage person-scene relations and propose a Global CNN model trained to predict positions and scales of heads directly from the full image. Second, we explicitly model pairwise relations among objects and train a Pairwise CNN model using a structured-output surrogate loss. The Local, Global and Pairwise models are combined into a joint CNN framework. To train and test our full model, we introduce a large dataset composed of 369,846 human heads annotated in 224,740 movie frames. We evaluate our method and demonstrate improvements of person head detection against several recent baselines in three datasets. We also show improvements of the detection speed provided by our model. This work has been published at ICCV 2015 [18]. The code and the new dataset developed in this work are available online at http://www.di.ens.fr/willow/research/headdetection/.

7.4.3. On Pairwise Costs for Network Flow Multi-Object Tracking

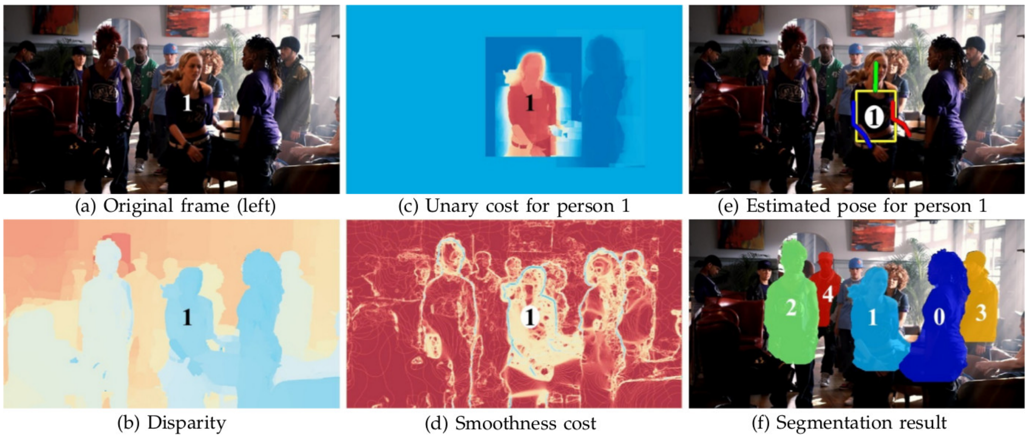
Participants: Visesh Chari, Simon Lacoste-Julien, Ivan Laptev, Josef Sivic.

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 published at CVPR 2015 [7].

7.4.4. Pose Estimation and Segmentation of Multiple People in Stereoscopic Movies

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

We describe a method to obtain a pixel-wise segmentation and pose estimation of multiple people in stereoscopic videos, illustrated in Figure 11. This task involves challenges such as dealing with unconstrained stereoscopic video, non-stationary cameras, and complex indoor and outdoor dynamic scenes with multiple people. We cast the problem as a discrete labelling task involving multiple person labels, devise a suitable cost function, and optimize it efficiently. The contributions of our work are two-fold: First, we develop a segmentation model incorporating person detections and learnt articulated pose segmentation masks, as well as colour, motion, and stereo disparity cues. The model also 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 587 annotated human poses, 1158 bounding box annotations and 686 pixel-wise segmentations 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.’s ACCV 2012). This work has been published at PAMI [4].



(a) Original frame (left)

(c) Unary cost for person 1

(e) Estimated pose for person 1

(b) Disparity

(d) Smoothness cost

(f) Segmentation result

Figure 11. Starting from a stereo pair (a), we estimate disparity maps (b). Using both appearance and disparity cues, we detect persons and estimate their poses (e). We combine pose information with disparity information and occlusion reasonning to compute the unary potentials of a CRF (c) and use standard color and motion cues to compute the binary terms (d). We optimize the CRF problem to produce the final, layered segmentation (f).

7.4.5. Weakly-Supervised Alignment of Video with Text

Participants: Piotr Bojanowski, Rémi Lajugie, Edouard Grave, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid.

In this work [6], we design a method for aligning natural language sentences with a video stream. Suppose that we are given a set of videos, along with natural language descriptions in the form of multiple sentences (e.g., manual annotations, movie scripts, sport summaries etc.), and that these sentences appear in the same temporal order as their visual counterparts. We propose here a method for aligning the two modalities, i.e., automatically providing a time stamp for every sentence (see Fig. 12). Given vectorial features for both video and text, we propose to cast this task as a temporal assignment problem, with an implicit linear mapping between the two feature modalities. We formulate this problem as an integer quadratic program, and solve its continuous convex relaxation using an efficient conditional gradient algorithm. Several rounding procedures are proposed to construct the final integer solution. After demonstrating significant improvements over the state of the art on the related task of aligning video with symbolic labels, we evaluate our method on a challenging dataset of videos with associated textual descriptions, using both bag-of-words and continuous representations for text. This work has been published at CVPR 2015 [6].

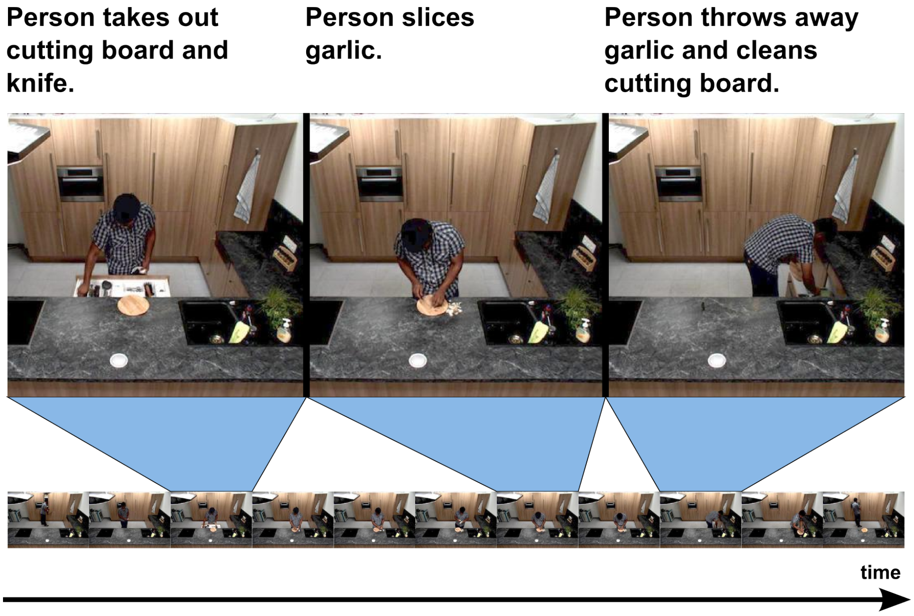

Person takes out cutting board and knife.
Person slices garlic.
Person throws away garlic and cleans cutting board.
time

Figure 12. Illustration of the text to video alignment problem. As an output, our model provides a temporal location for every sentence.

7.4.6. Unsupervised learning from narrated instruction videos

Participants: Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, Simon Lacoste-Julien.

In [20], we address the problem of automatically learning the main steps to complete a certain task, such as changing a car tire, from a set of narrated instruction videos. The contributions of this paper are three-fold. First, we develop a new unsupervised learning approach that takes advantage of the complementary nature of the input video and the associated narration. The method solves two clustering problems, one in text and one in video, applied one after each other and linked by joint constraints to obtain a single coherent sequence of steps in both modalities. Second, we collect and annotate a new challenging dataset of real-world instruction videos from the Internet. The dataset contains about 800,000 frames for five different tasks that include complex interactions between people and objects, and are captured in a variety of indoor and outdoor settings. Third, we experimentally demonstrate that the proposed method can automatically discover, in an unsupervised manner, the main steps to achieve the task and locate the steps in the input videos. This work is under review.

7.4.7. Long-term Temporal Convolutions for Action Recognition

Participants: Gül Varol, Ivan Laptev, Cordelia Schmid.

Typical human actions such as hand-shaking and drinking last several seconds and exhibit characteristic spatio-temporal structure. Recent methods attempt to capture this structure and learn action representations with convolutional neural networks. Such representations, however, are typically learned at the level of single frames or short video clips and fail to model actions at their full temporal scale. In [27], we learn video representations using neural networks with long-term temporal convolutions. We demonstrate that CNN models with increased temporal extents improve the accuracy of action recognition despite reduced spatial resolution. We also study the impact of different low-level representations, such as raw values of video pixels and optical flow vector fields and demonstrate the importance of high-quality optical flow estimation for learning accurate action models. We report state-of-the-art results on two challenging benchmarks for human action recognition UCF101 and HMDB51. This work is under review. The results for the proposed method are illustrated in Figure 13.

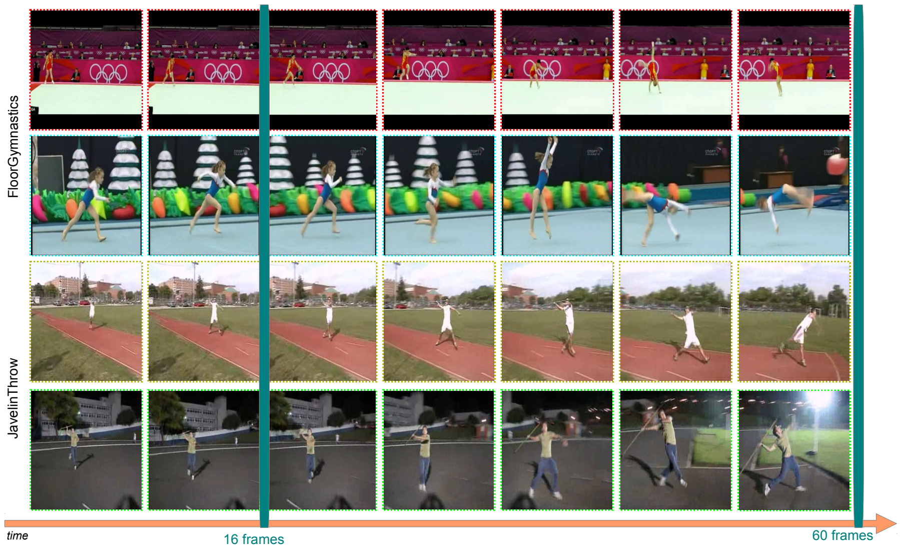


Figure 13. The highest improvement of long-term temporal convolutions in terms of class accuracy is for "JavelinThrow". For 16-frame network, it is mostly confused with "FloorGymnastics" class. We visualize sample videos with 7 frames extracted at every 8 frames. The intuitive explanation is that both classes start by running for a few seconds and then the actual action takes place. Long-term temporal convolutions with 60 frames can capture this interval, whereas 16-frame networks fail to recognize such long-term activities.

7.4.8. Thin-Slicing forPose: Learning to Understand Pose without Explicit Pose Estimation

Participants: Suha Kwak, Minsu Cho, Ivan Laptev.

In [23], we address the problem of learning a pose-aware, compact embedding that projects images with similar human poses to be placed close-by in the embedding space (Figure 14). The embedding function is built on a deep convolutional network, and trained with a triplet-based rank constraint on real image data. This architecture allows us to learn a robust representation that captures differences in human poses by effectively factoring out variations in clothing, background, and imaging conditions in the wild. For a variety of pose-related tasks, the proposed pose embedding provides a cost-efficient and natural alternative to explicit pose estimation, circumventing challenges of localizing body joints. We demonstrate the efficacy of the embedding on pose-based image retrieval and action recognition problems. This work is under review.

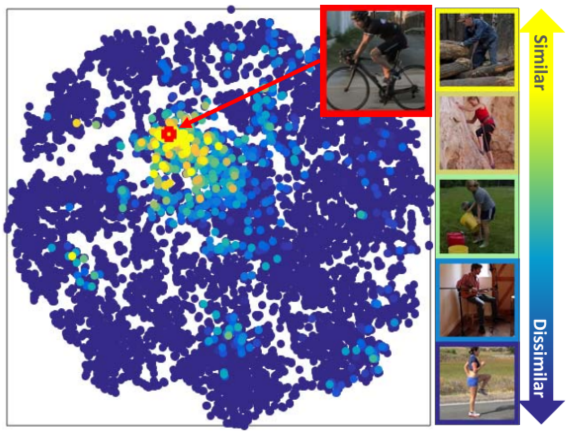


Figure 14. The manifold of our pose embedding visualized using t-SNE. Each point represents a human pose image. To better show correlation between the pose embedding and annotated pose, we color-code pose similarities in annotation between an arbitrary target image (red box) and all the other images. Selected examples of color-coded images are illustrated in the right-hand side. Images similar with the target in annotated pose are colored in yellow, otherwise in blue. As can be seen, yellow images lie closer by the target in general, which indicates that a position on the embedding space implicitly represents a human pose.

7.4.9. Instance-level video segmentation from object tracks

Participants: Guillaume Seguin, Piotr Bojanowski, Rémi Lajugie, Ivan Laptev.

In [26], we address the problem of segmenting multiple object instances in complex videos. Our method does not require manual pixel-level annotation for training, and relies instead on readily-available object detectors or visual object tracking only. Given object bounding boxes at input as shown in Figure 15, we cast video segmentation as a weakly-supervised learning problem. Our proposed objective combines (a) a discriminative clustering term for background segmentation, (b) a spectral clustering one for grouping pixels of same object instances, and (c) linear constraints enabling instance-level segmentation. We propose a convex relaxation of this problem and solve it efficiently using the Frank-Wolfe algorithm. We report results and compare our method to several baselines on a new video dataset for multi-instance person segmentation. This work is under review.

8. Bilateral Contracts and Grants with Industry

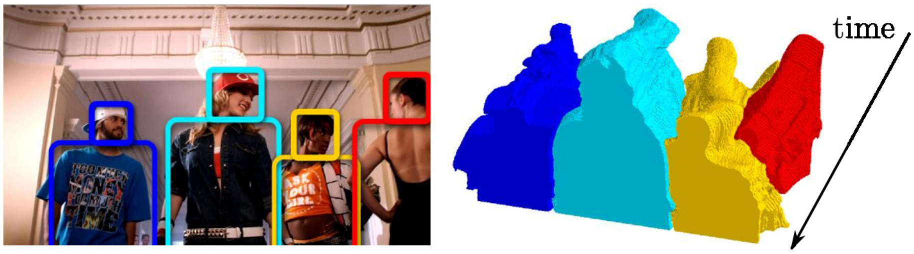


Figure 15. Results of our method applied to multi-person segmentation in a sample video from our database. Given an input video together with the tracks of object bounding boxes (left), our method finds pixel-wise segmentation for each object instance across video frames (right).

8.1. Facebook AI Research Paris: Weakly-supervised interpretation of image and video data (Inria)

Participants: Jean Ponce, Minsu Cho, Ivan Laptev, Josef Sivic.

We will develop in this project (Facebook gift) new models of image and video content, as well as new recognition architectures and algorithms, to address the problem of understanding the visual content of images and videos using weak forms of supervision, such as the fact that multiple images contain instances of the same objects, or the textual information available in television or film scripts.

8.2. 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.

8.3. Google: Structured learning from video and natural language (Inria)

Participants: Simon Lacoste-Julien, Ivan Laptev, Josef Sivic.

People can easily learn how to change a flat tire of a car or assemble an IKEA shelve by observing other people doing the same task, for example, by watching a narrated instruction video. In addition, they can easily perform the same task in a different context, for example, at their home. This involves advanced visual intelligence abilities such as recognition of objects and their function as well as interpreting sequences of human actions that achieve a specific task. However, currently there is no artificial system with a similar cognitive visual competence. The goal of this proposal is to develop models, representations and learning algorithms for automatic understanding of complex human activities from videos narrated with natural language.

8.4. MSR-Inria joint lab: Image and video mining for science and humanities (Inria)

Participants: Leon Bottou [Facebook], 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.

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. 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).

9.2. European Initiatives

9.2.1. 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 2016 by the ERC Advanced Grant "VideoWorld" awarded to Jean Ponce by the European Research Council.

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

9.2.2. 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?’, (classification) or sometimes ’Find the dog in this photo’ (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. 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?’. Solving such queries is the aim of this proposal. My goal 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 main 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. The project is timely as it builds upon the two key recent technological advances: (a) the immense progress in visual recognition of objects, scenes and human actions achieved in the last ten years, as well as (b) the emergence of a massive amount of public image and video data now available to train visual models. 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 detection of abnormal events, which are likely to revolutionise today’s approaches to crime protection, hazard prevention, elderly care, and many others.’

9.2.3. 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.’

9.2.4. 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).

9.3. International Initiatives

9.3.1. 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.

9.3.2. Inria CityLab initiative

Participants: Josef Sivic, Jean Ponce, Ivan Laptev, Alexei 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).

9.4. International Research Visitors

9.4.1. Visits of International Scientists

Prof. Alexei Efros (UC Berkeley) has visited Willow for one month in 2015. Prof. John Canny (UC Berkeley) has visited Willow during three months in 2015 within the framework of Inria’s International Chair program.

9.4.1.1. Internships

Filip Srajer (Czech Technical University) has been a visiting MSc student at Willow in Feb 2015. Nishant Agrawal (IIIT, India) has been a visiting intern at Willow for three months in 2015. Yumin Suh (Seoul National Univ., South Korea) has been a visiting intern at Willow for five months in 2015. Michail Nikita (Moscow State Univ., Russia) has been a visiting intern at Willow for three weeks in 2015.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific events organisation

10.1.1.1. General chair, scientific chair

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

10.1.1.2. Member of the organizing committees

- Workshop co-organizer, THUMOS Challenge 2015: Action Recognition with a Large Number of Classes, in conjunction with CVPR’15, Boston, USA (Ivan Laptev).

10.1.2. Scientific events selection

10.1.2.1. Area chairs

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015 (I. Laptev, J. Ponce).
- International Conference on Computer Vision (ICCV), 2015 (I. Laptev).

10.1.2.2. Member of the conference program committee

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015 (R. Arandjelović, P. Bojanowski, M. Cho, J. Sivic).
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 (R. Arandjelović, P. Bojanowski, G. Cheron, M. Cho, S. Kwak, J. Sivic).
- International Conference on Computer Vision (ICCV), 2015 (R. Arandjelović, P. Bojanowski, M. Cho, J. Sivic).
- ACM Multimedia 15 doctoral consortium (R. Arandjelović).
- 29th Conference on Artificial Intelligence (AAAI), 2015 (I. Laptev).
- International Conference on Learning Representations, 2015 (J. Sivic).

10.1.3. Journals

10.1.3.1. 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.

10.1.3.2. Reviewer

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

10.1.4. Others

- J. Sivic is senior fellow of the Neural Computation and Adaptive Perception program of the Canadian Institute of Advanced Research
- R. Arandjelović obtained the Best reviewer award at International Conference on Computer Vision (ICCV), 2015

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

- Master : J. Ponce, “Introduction to computer vision”, M1, 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.
- Doctorat : I. Laptev gave a tutorial at the 9th Russian Summer School in Information Retrieval (RuSSIR), St Peresburg, August 2015, 5h.

10.2.2. Supervision

PhD : Mathieu Aubry, “Visual recognition and retrieval of 3D objects and scenes”, graduated in 2015, J. Sivic and D. Cremers (TU Munich).

PhD : Vincent Delaitre, “Modeling and recognition of human-object interactions”, graduated in 2015, I. Laptev and J. Sivic.

PhD in progress : Gül Varol, “Deep learning methods for video interpretation”, starting in Oct 2015, I. Laptev, C. Schmid.

PhD in progress : Julia Peyre, “Learning to reason about scenes from images and language", starting in Oct 2015, C. Schmid, I. Laptev, J. Sivic.

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

PhD in progress : Rafael Sampaio de Rezende, started in 2013, 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 in Jan 2015, M. Aubry 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.

10.2.3. Juries

PhD thesis committee:

- Vincent Buso, Université Bordeaux I, France, 2015 (I. Laptev).
- Danila Potapov, Université Grenoble Alpes, France, 2015 (I. Laptev, rapporteur).
- Thomas Pfister, University of Oxford, UK, 2015, (I. Laptev, external examiner)
- Xi Chen, Aalto University, Finland, 2015 (I. Laptev, rapporteur).

Other:

Member of the PSL Research Council, 2012- (J. Ponce).

Member of Inria Commission de developpement technologique (CDT), 2012- (J. Sivic).

10.3. Invited presentations

- R. Arandjelović, Invited talk, DeepMind, London, United Kingdom, October 13, 2015
- M. Cho, Invited talk, Postech, Pohang, South Korea, November 30, 2015
- I. Laptev, Amazon, Seattle, USA, Nov, 2015.
- I. Laptev, Deep Video Workshop, Santa Cruz, USA, Nov, 2015.
- I. Laptev, EHESS, Paris, France, Nov 16 - 18, 2015.
- I. Laptev, Keynote Speaker, NCCV, Sept. 14-15, Lunteren, The Netherlands, 2015.
- I. Laptev, Institute for Computer Graphics and VisionTU Graz, Austria, July, 2015.
- I. Laptev, DALI Workshop, La Palma, Spain, April 2015.
- I. Laptev, Bauman Moscow State Technical University, Moscow, Russia, April 2015.
- I. Laptev, CVPR'15 Area Chair Workshop, Boston, MA, USA, March 2015.
- I. Laptev, GDR-ISIS, Paris, France, March, 2015.
- I. Laptev, Deep Learning Summit, San Francisco, USA, Jan. 2015.
- J. Ponce, IAS Workshop on Functoriality in Geometric Data, Hong Kong, April 2015.
- J. Ponce, ORASIS Workshop, Amiens, June 2015.
- J. Ponce, Optimization in Machine learning, vision and image processing Workshop, Toulouse, October 2015.
- J. Ponce, SAMSUNG DMC, Seoul, Korea, October 2015.
- J. Ponce, 12th International Conference on Ubiquitous Robots and Ambient Intelligence, Goyang City, Korea, October 2015.
- J. Ponce, International Conference in Big Data and Information Analytics, Xi'an, China, October 2015.
- J. Ponce, Xi'an Jiaotong University, Xi'an, China, October 2015.
- J. Ponce, Software and Digital Humanities Workshop, Paris, November 2015.
- J. Ponce, Algortithms for Human-Robot Interaction, Berkeley, USA, November 2015.
- J. Ponce, Distinguished speaker, Faculty of Engineering, Hong Kong University, December 2015.
- J. Sivic, Invited talk, the CIFAR workshop, Montreal, December 2015.
- J. Sivic, Invited talk, the Workshop on Visual Place Recognition in Changing Environments, CVPR 2015.
- J. Sivic, Invited talk, the Aristote seminar on scientific computing for smart cities, 2015.
- J. Sivic, Invited talk, the High Visual Computing seminar, 2015.
- J. Sivic, Talk at Inria ComDir, 2015.

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] M. AUBRY. *Representing 3D models for alignment and recognition*, ENS, May 2015, https://tel.archives-ouvertes.fr/tel-01160300

[2] V. DELAITRE. *Modeling and Recognizing Interactions between People, Objects and Scenes*, ENS Paris - Ecole Normale Supérieure de Paris, April 2015, https://hal.inria.fr/tel-01256076

Articles in International Peer-Reviewed Journals

[3] C. DOERSCH, S. SINGH, A. GUPTA, J. SIVIC, A. EFROS. *What Makes Paris Look Like Paris?*, in "Communications of the ACM", December 2015, vol. 58, n^o 12, pp. 103-110 [*DOI :* 10.1145/2830541], https://hal.inria.fr/hal-01248528

[4] 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", August 2015, vol. 37, n^o 8, pp. 1643 - 1655 [*DOI :* 10.1109/TPAMI.2014.2369050], https://hal.inria.fr/hal-01089660

[5] A. TORII, J. SIVIC, M. OKUTOMI, T. PAJDLA. *Visual Place Recognition with Repetitive Structures*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", March 2015, pp. 1-14 [*DOI :* 10.1109/TPAMI.2015.2409868], https://hal.inria.fr/hal-01152483

International Conferences with Proceedings

[6] P. BOJANOWSKI, R. LAJUGIE, E. GRAVE, F. BACH, I. LAPTEV, J. PONCE, C. SCHMID. *Weakly-Supervised Alignment of Video With Text*, in "ICCV 2015 - IEEE International Conference on Computer Vision", Santiago, Chile, December 2015, https://hal.inria.fr/hal-01154523

[7] V. CHARI, S. LACOSTE-JULIEN, I. LAPTEV, J. SIVIC. *On Pairwise Cost for Multi-Object Network Flow Tracking*, in "CVPR 2015 - 28th IEEE Conference on Computer Vision and Pattern Recognition", Boston, United States, June 2015, https://hal.inria.fr/hal-01110678

[8] M. CHO, S. KWAK, C. SCHMID, J. PONCE. *Unsupervised Object Discovery and Localization in the Wild: Part-based Matching with Bottom-up Region Proposals*, in "CVPR 2015 - IEEE Conference on Computer Vision & Pattern Recognition", Boston, United States, June 2015, https://hal.inria.fr/hal-01110036

[9] G. CHÉRON, I. LAPTEV, C. SCHMID. *P-CNN: Pose-based CNN Features for Action Recognition*, in "ICCV 2015 - IEEE International Conference on Computer Vision", Santiago, Chile, December 2015, https://hal.inria.fr/hal-01187690

[10] B. HAM, M. CHO, J. PONCE. *Robust Image Filtering Using Joint Static and Dynamic Guidance*, in "CVPR 2015", BOSTON, United States, June 2015, https://hal.inria.fr/hal-01240280

[11] S. KWAK, M. CHO, I. LAPTEV, J. PONCE, C. SCHMID. *Unsupervised Object Discovery and Tracking in Video Collections*, in "ICCV 2015 - IEEE International Conference on Computer Vision", Santiago, Chile, December 2015, https://hal.archives-ouvertes.fr/hal-01153017

[12] R. LAJUGIE, P. BOJANOWSKI, P. CUVILLIER, S. ARLOT, F. BACH. *A weakly-supervised discriminative model for audio-to-score alignment*, in "41st International Conference on Acoustics, Speech, and Signal Processing (ICASSP)", Shanghai, China, Proceedings of the 41st International Conference on Acoustics, Speech, and Signal Processing (ICASSP), March 2016, https://hal.archives-ouvertes.fr/hal-01251018

[13] S. LEE, N. MAISONNEUVE, D. CRANDALL, A. A. EFROS, J. SIVIC. *Linking Past to Present: Discovering Style in Two Centuries of Architecture*, in "IEEE International Conference on Computational Photography", Houston, United States, April 2015, https://hal.inria.fr/hal-01152482

[14] M. Oquab, L. Bottou, I. Laptev, J. Sivic. *Is object localization for free? – Weakly-supervised learning with convolutional neural networks*, in "IEEE Conference on Computer Vision and Pattern Recognition", Boston, United States, June 2015, https://hal.inria.fr/hal-01015140

[15] J. Sun, W. Cao, Z. Xu, J. Ponce. *Learning a convolutional neural network for non-uniform motion blur removal*, in "CVPR 2015 - IEEE Conference on Computer Vision and Pattern Recognition 2015", Boston, United States, IEEE, June 2015 [*DOI :* 10.1109/CVPR.2015.7298677], https://hal.inria.fr/hal-01250478

[16] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, T. Pajdla. *24/7 place recognition by view synthesis*, in "CVPR 2015 - 28th IEEE Conference on Computer Vision and Pattern Recognition", Boston, United States, June 2015, https://hal.inria.fr/hal-01147212

[17] M. Trager, M. Hebert, J. Ponce. *The joint image handbook*, in "ICCV 2015", Santiago, Chile, December 2015, https://hal.archives-ouvertes.fr/hal-01249171

[18] T.-H. Vu, A. Osokin, I. Laptev. *Context-aware CNNs for person head detection*, in "ICCV 2015 - IEEE International Conference on Computer Vision", Santiago, Chile, December 2015, To appear in International Conference on Computer Vision (ICCV), 2015, https://hal.archives-ouvertes.fr/hal-01237618

Scientific Books (or Scientific Book chapters)

[19] M. Aubry, B. C. Russell, J. Sivic. *Visual geo-localization of non-photographic depictions via 2D-3D alignment*, in "Visual Analysis and Geolocalization of Large-Scale Imagery", Springer (editor), 2015, https://hal.archives-ouvertes.fr/hal-01119203

Other Publications

[20] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, S. Lacoste-Julien. *Unsupervised Learning from Narrated Instruction Videos*, June 2015, improved NLP method and bigger dataset, https://hal.inria.fr/hal-01171193

[21] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, J. Sivic. *NetVLAD: CNN architecture for weakly supervised place recognition*, November 2015, working paper or preprint, https://hal.inria.fr/hal-01242052

[22] B. Ham, M. Cho, C. Schmid, J. Ponce. *Proposal Flow*, December 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01240281

[23] S. Kwak, M. Cho, I. Laptev. *Thin-Slicing for Pose: Learning to Understand Pose without Explicit Pose Estimation*, December 2015, working paper or preprint, https://hal.inria.fr/hal-01242724

[24] R. Lajugie, P. Bojanowski, S. Arlot, F. Bach. *Semidefinite and Spectral Relaxations for Multi-Label Classification*, June 2015, working paper or preprint, https://hal.inria.fr/hal-01159321

[25] J. Ponce, M. Hebert, M. Trager. *Trinocular Geometry Revisited*, 2015, Submitted to International Journal of Computer Vision, https://hal.archives-ouvertes.fr/hal-01152348

[26] G. Seguin, P. Bojanowski, R. Lajugie, I. Laptev. *Instance-level video segmentation from object tracks*, January 2016, working paper or preprint, https://hal.inria.fr/hal-01255765

[27] G. VAROL, I. LAPTEV, C. SCHMID. *Long-term Temporal Convolutions for Action Recognition*, December 2015, working paper or preprint, https://hal.inria.fr/hal-01241518