IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2016

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

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

Minsu Cho [Inria, Starting Research position, until Aug 2016]
Ivan Laptev [Inria, Senior Researcher, HDR]
Josef Sivic [Inria, Senior Researcher, HDR]

Faculty Member

Jean Ponce [Team leader, ENS Paris, Professor]

Engineers

Jonathan Chemla [Inria, until Aug 2016]
Petr Gronat [Inria]
Antony Marion [Inria, until Mar 2016]
Ignacio Rocco Spremolla [Inria, from Apr 2016]

PhD Students

Guilhem Cheron [Inria]
Theophile Dalens [Inria]
Vadim Kantorov [Inria]
Antoine Miech [Inria, from Oct 2016]
Maxime Oquab [Inria]
Julia Peyre [Inria]
Rafael Sampaio de Rezende [Inria]
Guillaume Seguin [ENS Paris, until Aug 2016]
Matthew Trager [Inria]
Gul Varol Simsekli [Inria]
Piotr Bojanowski [Inria, until Mar 2016]
Tuan Hung Vu [Inria]

Post-Doctoral Fellows

Anton Osokin [Inria]
Relja Arandjelovic [Inria, until Jul 2016]

Andrei Bursuc [Inria, until Oct 2016]
Bumsub Ham [Inria, until Jul 2016]
Suha Kwak [Inria, until Mar 2016]

Visiting Scientists

John Canny [UC Berkeley, Professor, Inria International Chair]
Alexei Efros [UC Berkeley, May-Jun 2016]
Sergiu Irimie [Inria, until Aug 2016]
Phillip Isola [UC Berkeley, Jun 2016]
Oleh Rybkin [Inria, Sep 2016]
Richard Zhang [UC Berkeley, Jun 2016]

Administrative Assistants

David Dinis [Inria, until Apr 2016]
Sarah Le [Inria, from Jul 2016]

Others

Kai Han [Inria]
Mathieu Aubry [ENPC]
Gunnar Atli Sigurdsson [Carnegie Mellon University]
Pavel Trutman [Czech Technical University]

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 two new Phd students: Antoine Miech (Inria) and Ignacio Rocco (inria). Alexei Efros (Professor, UC Berkeley, USA) visited Willow during May-June with his postdoc Phillip Isola and Phd student Richard Zhang. John Canny (Professor, UC Berkeley, USA) visited 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 are outlined in detail in Section. 7.1.

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 in this area is outlined in detail in Section 7.2.

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

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

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 is outlined in detail in Section 7.3.

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.

- **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.
- **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 various tasks such as human action recognition, person detection, segmentation and tracking.

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

5.1.1. Awards

- Jean Ponce (together with Svetlana Lazebnik and Cordelia Schmid) received the Longuet-Higgins Prize for “Fundamental contributions in Computer Vision", awarded at the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

6. New Software and Platforms

6.1. 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" [9]. 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/.

6.2. Unsupervised learning from narrated instruction videos

Open source release of the software package for our paper "Unsupervised learning from narrated instruction videos" . It provides a full implementation of the method, including code for weakly supervised training from instruction video, as well as trained models. Links to all of these are available at our project page http://www.di.ens.fr/willow/research/instructionvideos/.

6.3. ContextLocNet: Context-aware deep network models for weakly supervised localization

Open source release of code reproducing the results in our "ContextLocNet: Context-aware deep network models for weakly supervised localization" [11]. It provides code for training models, testing on standard datasets and trained models. It can be found online at https://github.com/vadimkantorov/contextlocnet.

6.4. Long-term Temporal Convolutions for Action Recognition

Open source release of the software package for our paper "Long-term Temporal Convolutions for Action Recognition" [20]. It provides code for training models, testing on standard datasets and trained models. Links are available at our project page http://www.di.ens.fr/willow/research/ltc/.

7. New Results

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

7.1.1. 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 published in IJCV [7] and example results are shown in figure 1.

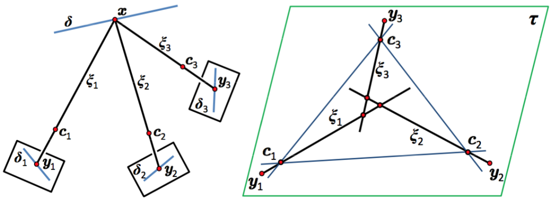

Figure 1. Left: Visual rays associated with three (correct) correspondences. Right: Degenerate epipolar constraints associated with three coplanar, but non-intersecting rays lying in the trifocal plane.

7.1.2. Consistency of silhouettes and their duals

Participants: Matthew Trager, Martial Hebert, Jean Ponce.

Silhouettes provide rich information on three-dimensional shape, since the intersection of the associated visual cones generates the "visual hull", which encloses and approximates the original shape. However, not all silhouettes can actually be projections of the same object in space: this simple observation has implications in object recognition and multi-view segmentation, and has been (often implicitly) used as a basis for camera calibration. In this paper, we investigate the conditions for multiple silhouettes, or more generally arbitrary closed image sets, to be geometrically "consistent". We present this notion as a natural generalization of traditional multi-view geometry, which deals with consistency for points. After discussing some general results, we present a "dual" formulation for consistency, that gives conditions for a family of planar sets to be sections of the same object. Finally, we introduce a more general notion of silhouette "compatibility" under partial knowledge of the camera projections, and point out some possible directions for future research. This work has been published in [16] and example results are shown in 2.

7.1.3. Congruences and Concurrent Lines in Multi-View Geometry

Participants: Jean Ponce, Bernd Sturmfels, Matthew Trager.

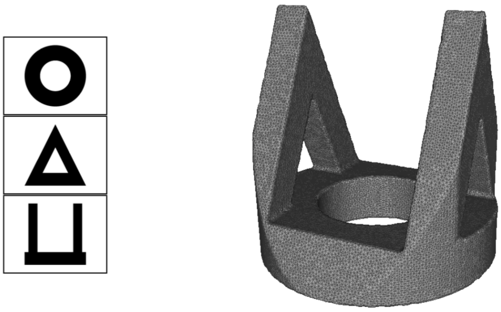

Figure 2. Geometrically consistent silhouettes are feasible projections of a single object.

We present a new framework for multi-view geometry in computer vision. A camera is a mapping between P^3 and a line congruence. This model, which ignores image planes and measurements, is a natural abstraction of traditional pinhole cameras. It includes two-slit cameras, pushbroom cameras, catadioptric cameras, and many more. We study the concurrent lines variety, which consists of n-tuples of lines in P^3 that intersect at a point. Combining its equations with those of various congruences, we derive constraints for corresponding images in multiple views. We also study photographic cameras which use image measurements and are modeled as rational maps from P^3 to P^2 or $P^1 \times P^1$. This work has been accepted for publication in [19] and example results are shown in 3.

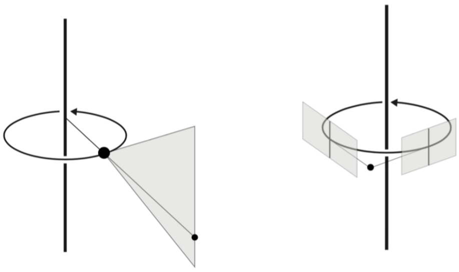

Figure 3. Non-central panoramic (left) and stereo panoramic cameras (right) are examples of non-linear cameras that can be modeled using line congruences.

7.1.4. NetVLAD: CNN architecture for weakly supervised place recognition

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

In [9], 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 has been published at CVPR 2016 [9]. Figure 4 shows some qualitative results.

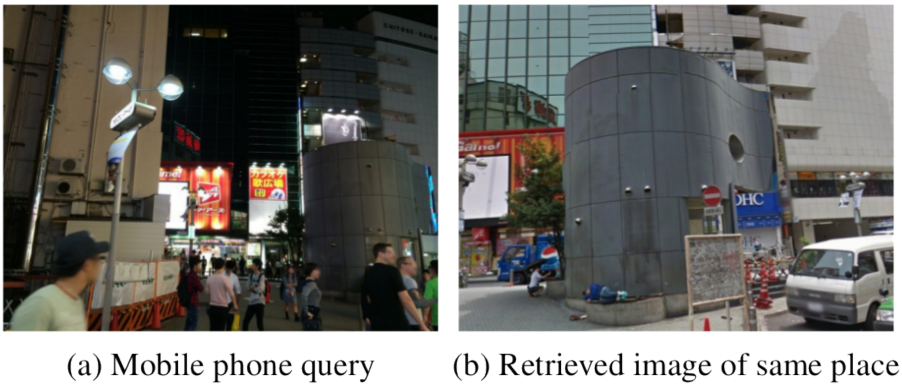

Figure 4. Our trained NetVLAD descriptor correctly recognizes the location (b) of the query photograph (a) despite the large amount of clutter (people, cars), changes in viewpoint and completely different illumination (night vs daytime).

7.1.5. Pairwise Quantization

Participants: Artem Babenko, Relja Arandjelović, Victor Lempitsky.

We consider the task of lossy compression of high-dimensional vectors through quantization. We propose the approach that learns quantization parameters by minimizing the distortion of scalar products and squared distances between pairs of points. This is in contrast to previous works that obtain these parameters through the minimization of the reconstruction error of individual points. The proposed approach proceeds by finding a linear transformation of the data that effectively reduces the minimization of the pairwise distortions to the minimization of individual reconstruction errors. After such transformation, any of the previously-proposed quantization approaches can be used. Despite the simplicity of this transformation, the experiments demonstrate that it achieves considerable reduction of the pairwise distortions compared to applying quantization directly to the untransformed data. This work has been published on arXiv [18] and submitted to Neurocomputing journal.

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

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

The aim of this work is to localize a query photograph by finding other images depicting the same place in a large geotagged image database. This is a challenging task due to changes in viewpoint, imaging conditions and the large size of the image database. The contribution of this work is two-fold. First, we cast the place recognition problem as a classification task and use the available geotags to train a classifier for each location in the database in a similar manner to per-exemplar SVMs in object recognition. Second, as only few positive training examples are available for each location, we propose a new approach to calibrate all the per-location SVM classifiers using *only* the negative examples. The calibration we propose relies on a significance measure essentially equivalent to the p-values classically used in statistical hypothesis testing. Experiments are performed on a database of 25,000 geotagged street view images of Pittsburgh and demonstrate improved place recognition accuracy of the proposed approach over the previous work. This work has been published at CVPR 2013, and a revised version that includes additional experimental results has been published at IJCV [3].

7.2. Category-level object and scene recognition

7.2.1. 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 [10], 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 has been published at CVPR 2016 [10]. The proposed method and its qualitative result are illustrated in Figure 5.

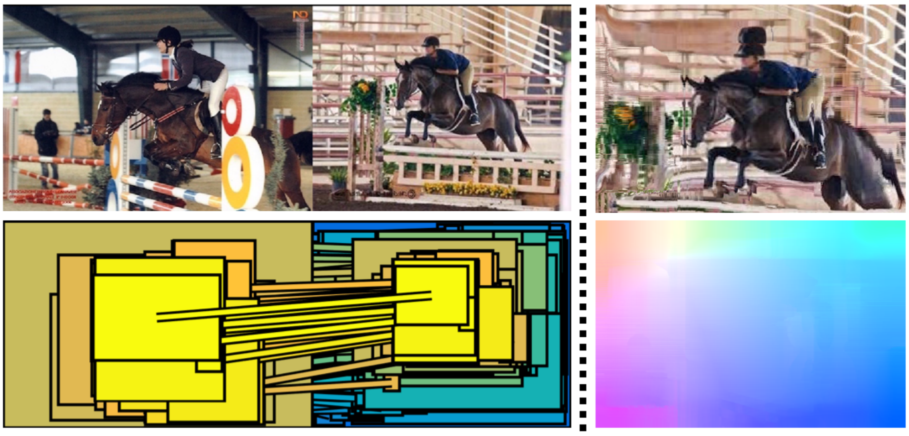

Figure 5. Proposal flow generates a reliable scene flow between similar images by establishing geometrically consistent correspondences between object proposals. (Left) Region-based scene flow by matching object proposals. (Right) Color-coded dense flow field generated from the region matches, and image warping using the flow.

7.2.1.1. Learning Discriminative Part Detectors for Image Classification and Cosegmentation

Participants: Jian Sun, Jean Ponce.

In this work, we address the problem of learning discriminative part detectors from image sets with category labels. We propose a novel latent SVM model regularized by group sparsity to learn these part detectors. Starting from a large set of initial parts, the group sparsity regularizer forces the model to jointly select and optimize a set of discriminative part detectors in a max-margin framework. We propose a stochastic version of a proximal algorithm to solve the corresponding optimization problem. We apply the proposed method to image classification and cosegmentation, and quantitative experiments with standard bench- marks show that it matches or improves upon the state of the art. The first version of this work has appeared at CVPR 2013. An extended version has been published at IJCV [6].

7.2.2. ContextLocNet: Context-aware deep network models for weakly supervised localization

Participants: Vadim Kantorov, Maxime Oquab, Minsu Cho, Ivan Laptev.

In [11] we aim to localize objects in images using image-level supervision only. Previous approaches to this problem mainly focus on discriminative object regions and often fail to locate precise object boundaries. In [11] we address this problem by introducing two types of context-aware guidance models, additive and contrastive models, that leverage their surrounding context regions to improve localization. The additive model encourages the predicted object region to be supported by its surrounding context region. The contrastive model encourages the predicted object region to be outstanding from its surrounding context region. Our approach benefits from the recent success of convolutional neural networks for object recognition and extends Fast R-CNN to weakly supervised object localization. Extensive experimental evaluation on the PASCAL VOC 2007 and 2012 benchmarks shows hat our context-aware approach significantly improves weakly supervised localization and detection. A high-level architecture of our model is presented in Figure 6, the project webpage is at http://www.di.ens.fr/willow/research/contextlocnet/.

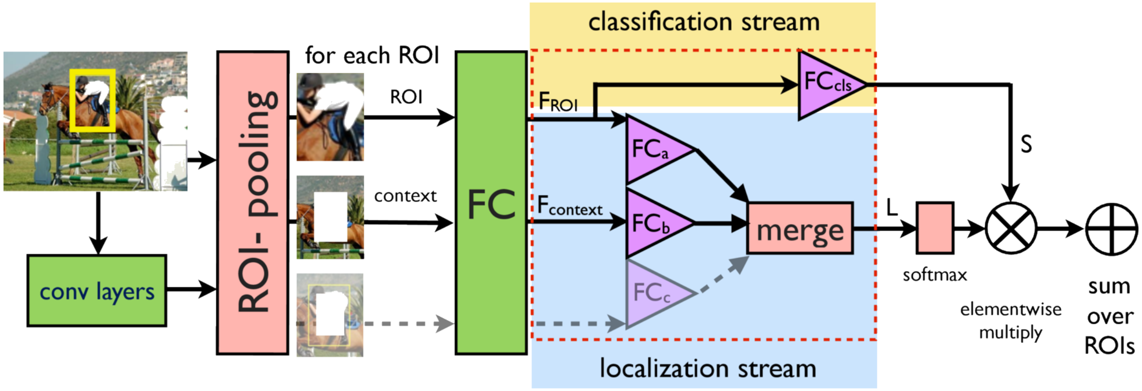

Figure 6. ContextLocNet improves localization by comparing an object score between a proposal and its context.

7.2.3. Faces In Places: Compound query retrieval

Participants: Yujie Zhong, Relja Arandjelović, Andrew Zisserman.

The goal of this work is to retrieve images containing both a target person and a target scene type from a large dataset of images. At run time this compound query is handled using a face classifier trained for the person, and an image classifier trained for the scene type. We make three contributions: first, we propose a hybrid convolutional neural network architecture that produces place-descriptors that are aware of faces and their corresponding descriptors. The network is trained to correctly classify a combination of face and scene classifier scores. Second, we propose an image synthesis system to render high quality fully-labelled face-and-place images, and train the network only from these synthetic images. Last, but not least, we collect and annotate a dataset of real images containing celebrities in different places, and use this dataset to evaluate the retrieval system. We demonstrate significantly improved retrieval performance for compound queries using the new face-aware place-descriptors. This work has been published at BMVC 2016 [17]. Figure 7 shows some qualitative results.

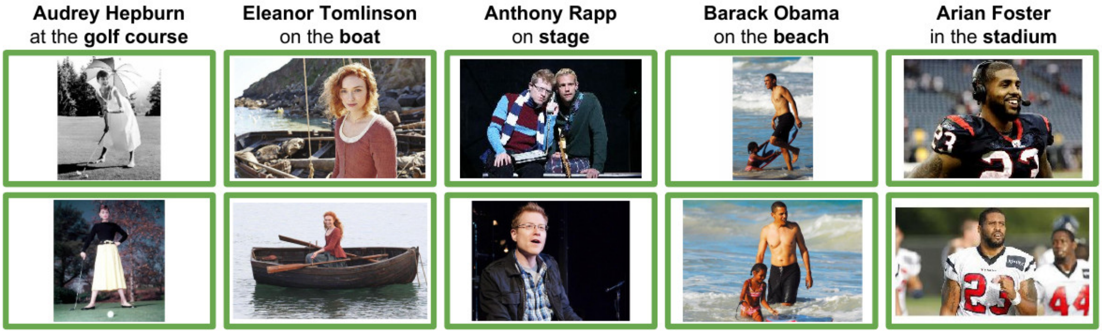

Figure 7. Examples of the top two retrieved images for various compound queries.

7.3. Image restoration, manipulation and enhancement

7.3.1. Robust Guided Image Filtering Using Nonconvex Potentials

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. A new revised version is currently in submission [4]. The SD filter is illustrated in Figure 8.

7.4. Human activity capture and classification

7.4.1. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding

Participants: Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, Abhinav Gupta.

Computer vision has a great potential to help our daily lives by searching for lost keys, watering flowers or reminding us to take a pill. To succeed with such tasks, computer vision methods need to be trained from real and diverse examples of our daily dynamic scenes. While most of such scenes are not particularly exciting, they typically do not appear on YouTube, in movies or TV broadcasts. So how do we collect sufficiently many diverse but boring samples representing our lives? We propose a novel Hollywood in Homes approach to collect such data. Instead of shooting videos in the lab, we ensure diversity by distributing and crowdsourcing the whole process of video creation from script writing to video recording and annotation. Following this procedure we collect a new dataset, *Charades*, with hundreds of people recording videos in their own homes, acting out casual everyday activities (see Figure 9). The dataset is composed of 9,848 annotated videos with an

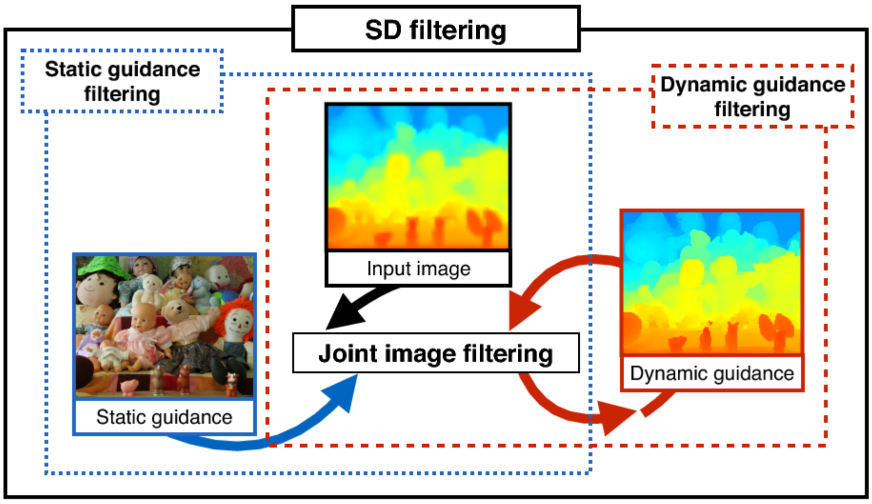

Figure 8. Sketch of joint image filtering and SD filtering: Static guidance filtering convolves an input image with a weight function computed from static guidance, as in the dotted blue box. Dynamic guidance filtering uses weight functions that are repeatedly obtained from regularized input images, 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.

average length of 30 seconds, showing activities of 267 people from three continents. Each video is annotated by multiple free-text descriptions, action labels, action intervals and classes of interacted objects. In total, Charades provides 27,847 video descriptions, 66,500 temporally localized intervals for 157 action classes and 41,104 labels for 46 object classes. Using this rich data, we evaluate and provide baseline results for several tasks including action recognition and automatic description generation. We believe that the realism, diversity, and casual nature of this dataset will present unique challenges and new opportunities for computer vision community. This work has been published at ECCV 2016 [15].

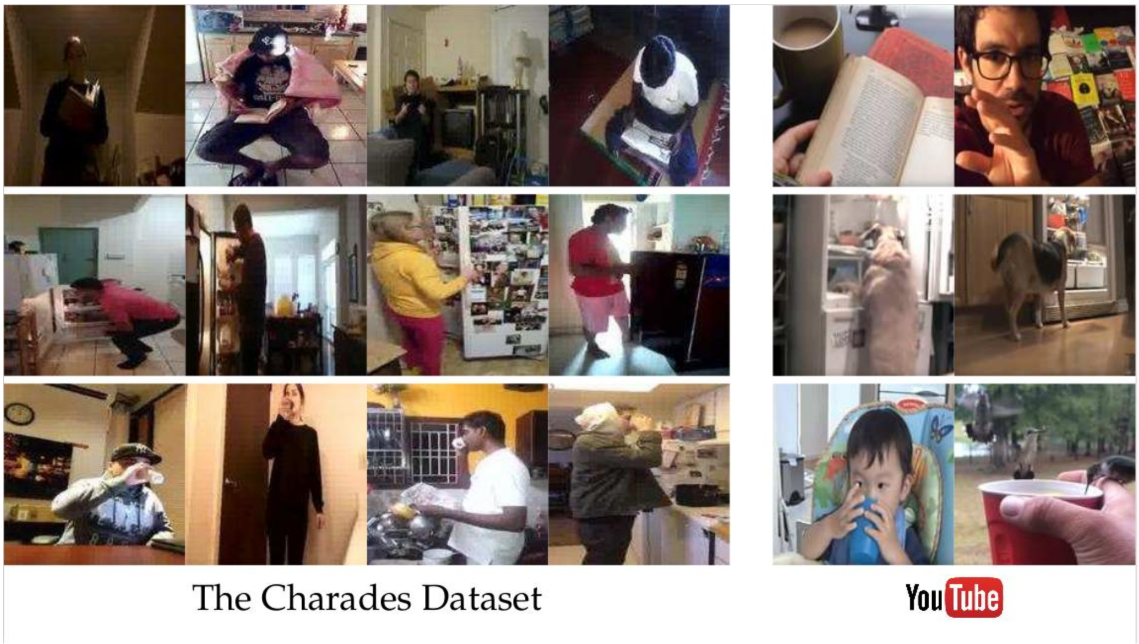

Figure 9. Comparison of actions in the Charades dataset and on YouTube: Reading a book, Opening a refrigerator, Drinking from a cup. YouTube returns entertaining and often atypical videos, while Charades contains typical everyday videos.

7.4.2. Unsupervised learning from narrated instruction videos

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

In [8], 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 has been published at CVPR 2016 [8].

7.4.3. Long-term Temporal Convolutions for Action Recognition

Participants: Gul 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 [20], 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 10.

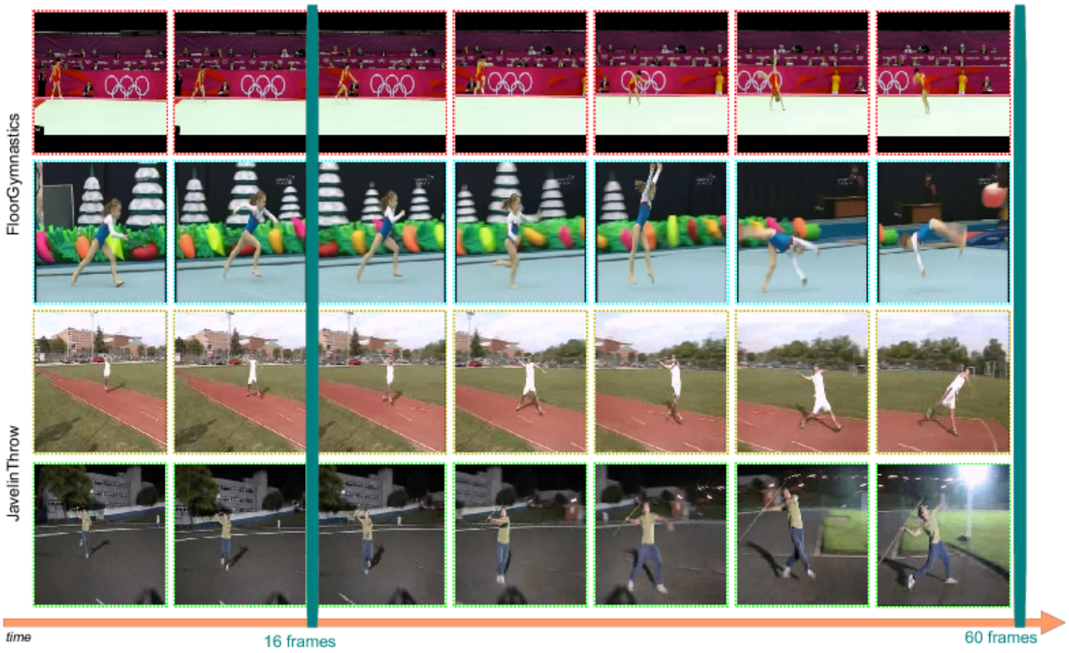

Figure 10. The highest improvement of long-term temporal convolutions in terms of class accuracy is for "JavelinThrow". For 16-frame network, it is mostly confused with "FloorGymnastics" class. We visualize sample videos with 7 frames extracted at every 8 frames. The intuitive explanation is that both 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.

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

Participants: Suha Kwak, Minsu Cho, Ivan Laptev.

In [12], 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 11). 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 has been published at CVPR 2016 [12].

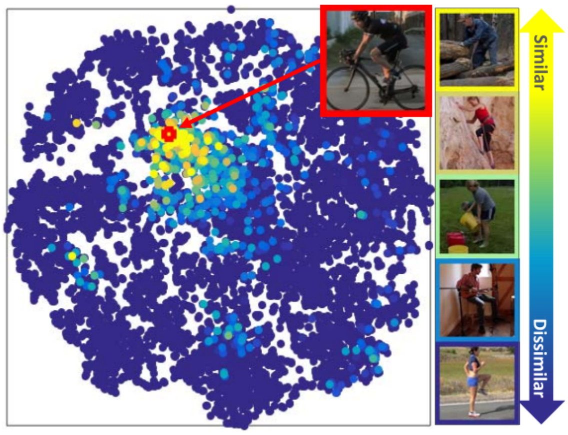

Figure 11. The manifold of our pose embedding visualized using t-SNE. Each point represents a human pose image. To better show correlation between the pose embedding and annotated pose, we color-code pose similarities in annotation between an arbitrary target image (red box) and all the other images. 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.

7.4.5. Instance-level video segmentation from object tracks

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

In [14], 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 12, 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 has been published at CVPR 2016.

8. Bilateral Contracts and Grants with Industry

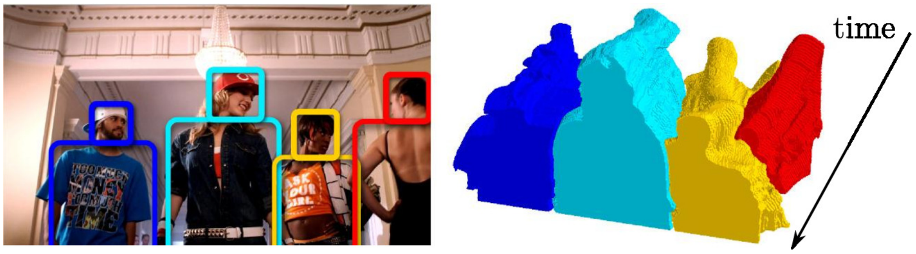

Figure 12. Results of our method applied to multi-person segmentation in a sample video from our database. Given an input video together with the tracks of object bounding boxes (left), our method finds pixel-wise segmentation for each object instance across video frames (right).

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.’

9.3. 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, USA) visited Willow during May-June with his postdoc Phillip Isola and Phd student Richard Zhang. Prof. John Canny (UC Berkeley) has visited Willow in 2016 within the framework of Inria’s International Chair program.

9.4.1.1. Internships

P. Trutman and O. Rybkin have visited Willow from Czech Technical University in Prague.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. General Chair, Scientific Chair

- I. Laptev will be program co-chair of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

10.1.1.2. Member of the Organizing Committees

- M. Trager is an organizer of "Minisymposium" on "Algebraic Vision" at the SIAM conference on Applied Algebraic Geometry (Atlanta, July 31st-August 4th 2017).

10.1.2. Scientific Events Selection

10.1.2.1. Area chairs

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 (I. Laptev).
- Asian Conference on Computer Vision (ACCV), 2016 (I. Laptev).
- International Conference on Computer Vision (ICCV), 2017 (J. Sivic).

10.1.2.2. Member of the Conference Program Committees

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 (R. Arandjelovic, A. Bursuc, P. Bojanowski, M. Cho, J. Sivic, G. Cheron).
- European Conference on Computer Vision (ECCV), 2016 (R. Arandjelovic, A. Bursuc, P. Bojanowski, G. Cheron, M. Cho, S. Kwak, J. Sivic, G. Cheron, I. Laptev).
- International Conference on Learning Representations, 2016 (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, J. Sivic).
- Foundations and Trends in Computer Graphics and Vision (J. Ponce).
- 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, G. Cheron, R. Arandjelovic).
- IEEE Transactions on Pattern Analysis and Machine Intelligence (R. Arandjelovic, P. Bojanowski, M. Cho, S. Kwak, G. Cheron, A. Bursuc).
- 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, A. Bursuc).
- 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. Arandjelovic and J. Sivic obtained the outstanding reviewer award at IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

10.1.5. Invited Talks

- V. Kantorov, Speaker at Second Christmas Colloquium on Computer Vision (December 2016, Moscow, SkolTech).
- I. Laptev, Invited talk, MailRu, Moscow, May 2016.
- I. Laptev, Invited talk, Skolkovo Robotics, Moscow, May 2016.
- I. Laptev, Invited talk, Deep Machine Intelligence and its Applications, SkolTech, Moscow, June 2016.
- I. Laptev, Invited talk, MSR-Inria Offsite Meeting, Paris, September, 2016.
- I. Laptev, Invited talk, University of Central Florida, Orlando, September, 2016.
- I. Laptev, Invited talk, Georgia Institute of Technology, Atlanta, September, 2016.
- I. Laptev, Invited talk, ECCV'16 Workshop on Brave new ideas for motion representations in videos, Amsterdam, October, 2016.
- I. Laptev, Invited talk, Open Day AI Innovation Factory , December, 2016.
- J. Ponce, Invited talk, New York University, January 2016.
- J. Ponce, Invited talk, Université Marne la Vallée, Mars 2016.
- J. Ponce, Invited talk, Workshop on Algebraic Vision, San Jose, May 2016.
- J. Ponce, Invited talk, Colloque LORIA, Nancy, May 2016.
- J. Ponce, Invited talk, Parthenos Workshop, Bordeaux, November 2016.

- J. Sivic, seminar, UC Berkeley, May, December, 2016.
- J. Sivic, invited talk, Brno University of Technology, April 2016.
- J. Sivic, Invited talk, the CIFAR workshop, Barcelona, December 2016.
- J. Sivic, Invited talk, Colloquium on Perspectives and New Challenges in Data Science, Ecole de Ponts ParisTech, 2016.
- M. Trager, invited speaker, AIM workshop “Algebraic Vision" (San Jose, May 2-6, 2016).

10.1.6. Leadership within the Scientific Community

- Member, advisory board, IBM Watson AI Xprize (J. Ponce).
- Member, steering committee, "France Intelligence Artificielle" initiative (J. Ponce).
- Member, advisory board, Computer Vision Foundation (J. Sivic).

10.1.7. Scientific Expertise

- J. Sivic gave an overview of state-of-the-art in computer vision at the seminar on artificial intelligence, Direction Generale des Entreprises (DGE) du Ministere de l’Economie, de l’Industrie et du Numerique, September, 2016.

10.1.8. Research Administration

- Member, Bureau du comité des projets, Inria, Paris (J. Ponce)
- Director, Department of Computer Science, Ecole normale supérieure (J. Ponce)
- Member, Scientific academic council, PSL Research University (J. Ponce)
- Member, Research representative committee, PSL Research University (J. Ponce).
- Member of Inria Commission de developpement technologique (CDT), 2012- (J. Sivic).
- Member of ANR evaluation committee (I. Laptev).

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

- Master : J. Ponce, "Introduction to computer vision", M1, Ecole normale superieure, 36h.
- Master : I. Laptev, J. Ponce and J. Sivic (together with C. Schmid, Inria Grenoble), "Object recognition and computer vision", M2, Ecole normale superieure, and MVA, Ecole normale superieure de Cachan, 36h.
- Master : I. Laptev, J. Ponce and J. Sivic, Cours PSL-ITI - Informatique, mathematiques appliques pour le traitement du signal et l’imagerie, 20h.

10.2.2. Supervision

PhD : Piotr Bojanowski, "Learning to annotate dynamic video scenes", graduated in 2016, I. Laptev, J. Ponce, C. Schmid and J. Sivic.

PhD : Guillaume Seguin, "Person analysis in stereoscopic movies", graduated in 2016, I. Laptev and J. Sivic.

PhD in progress : Ignacio Rocco, “Estimating correspondence between images via convolutional neural networks”, started in Jan 2017, J. Sivic, R. Arandjelovic (Google DeepMind).

PhD in progress : Antoine Miech, “Understanding long-term temporal structure of videos Phd thesis proposal", started in Oct 2016, I. Laptev, J. Sivic, P. Bojanowski (Facebook AI Research).

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

PhD in progress : Julia Peyre, “Learning to reason about scenes from images and language", started 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 (Inria SIERRA / U. Montreal).

PhD in progress : Rafael Sampaio de Rezende, started in 2013, J.Ponce.

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

PhD in progress : Theophile 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 (Facebook AI Research), 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:
  - Stavros Tsogkas, Ecole Centrale, France, 2016 (J. Sivic, examinateur).
  - Sesh Karri, Ecole Normale Superieure, France, 2016 (J. Sivic, examinateur).
  - Elliot Crowley, University of Oxford, UK, 2016, (J. Sivic, external examiner)
  - Olivier Frigo, Universite Paris Descartes, France, 2016 (J. Sivic, rapporteur).
  - Mattis Paulin, Inria Grenoble, France, 2017 (J. Sivic, rapporteur).
  - Francesco Massa, ENPC, France 2017 (J. Sivic, examinateur).
  - Philippe Weinzaepfel, Universite Grenoble Alpes, France, 2015 (I. Laptev, rapporteur).
  - Guillaume Seguin, Ecole Normale Superieure, France, 2016 (I. Laptev, J.Ponce, J. Sivic, examinateurs).
  - Piotr Bojanowski, Ecole Normale Superieure, France, 2016 (I. Laptev, J.Ponce, J. Sivic, examinateurs).
  - Ala Aboudib, Télécom Bretagne, France, 2016 (J. Ponce).
  - Philippe Weinzaepfel, Universite Grenoble Alpes, France, 2016 (J. Ponce).

10.3. Popularization

- Participation to the round table on "L'IA est-elle réservée aux GAFA", NUMA, June 2016 (J. Ponce).
- Participation to the round table on "Fictions, magie numerique et realites", Post-digital program, ENS/PSL Research University, October 2016 (J. Ponce).
- Debate with Jacques Attali, "Intelligence Artificielle, science avec conscience?", "Intelligence Artificielle : de la technique au business" Conference, December 2016 (J. Ponce).
- Participation to the round table on AI, Liberté Living Lab, December 2016 (J. Ponce).
- Participation to the round table on ethics at the Senate's public hearing on Artificial Intelligence, January 2017 (J. Ponce).
- Interview on Nolife 56Kast (https://www.youtube.com/watch?v=8UgH8_J2ugU) (J. Ponce).
- Interview in Le Monde (http://www.lemonde.fr/pixels/article/2016/01/08/intelligence-artificielle-ce-que-voient-les-machines_4843858_4408996.html) (J. Ponce).

- Interview in Télérama (http://www.telerama.fr/monde/trouver-le-calme-reconstituer-palmyre-ou-choisir-un-traitement-grace-a-l-ia,141131.php) (J. Ponce).

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] P. Bojanowski. *Learning to annotate dynamic video scenes*, Ecole normale supérieure, June 2016, https://hal.inria.fr/tel-01364560

[2] G. Seguin. *Person analysis in stereoscopic movies*, Ecole normale superieure, April 2016, https://tel.archives-ouvertes.fr/tel-01311143

Articles in International Peer-Reviewed Journals

[3] P. Gronát, G. Obozinski, J. Sivic, T. Pajdla. *Learning and calibrating per-location classifiers for visual place recognition*, in "International Journal of Computer Vision", April 2016, vol. 118, n^{O} 3, pp. 319-336 [*DOI :* 10.1007/s11263-015-0878-x], https://hal.inria.fr/hal-01418239

[4] B. Ham, M. Cho, J. Ponce. *Robust Guided Image Filtering Using Nonconvex Potentials*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2017, Accepted pending minor revision, https://hal.archives-ouvertes.fr/hal-01279857

[5] H. R. Idrees, A. R. Zamir, Y.-G. Jiang, A. R. Gorban, I. R. Laptev, R. R. Sukthankar, M. R. Shah. *The THUMOS challenge on action recognition for videos "in the wild"*, in "Computer Vision and Image Understanding", 2016 [*DOI :* 10.1016/j.cviu.2016.10.018], https://hal.inria.fr/hal-01431525

[6] J. Sun, J. Ponce. *Learning Dictionary of Discriminative Part Detectors for Image Categorization and Cosegmentation*, in "International Journal of Computer Vision", March 2016 [*DOI :* 10.1007/s11263-016-0899-0], https://hal.archives-ouvertes.fr/hal-01064637

[7] M. Trager, J. Ponce, M. Hebert. *Trinocular Geometry Revisited*, in "International Journal on Computer Vision (IJCV)", 2016, https://hal.archives-ouvertes.fr/hal-01152348

International Conferences with Proceedings

[8] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, S. Lacoste-Julien. *Unsupervised Learning from Narrated Instruction Videos*, in "CVPR2016 - 29th IEEE Conference on Computer Vision and Pattern Recognition", Las Vegas, United States, June 2016, https://hal.inria.fr/hal-01171193

[9] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, J. Sivic. *NetVLAD: CNN architecture for weakly supervised place recognition*, in "CVPR 2016 - 29th IEEE Conference on Computer Vision and Pattern Recognition", Las Vegas, United States, June 2016, https://hal.inria.fr/hal-01242052

[10] B. Ham, M. Cho, C. Schmid, J. Ponce. *Proposal Flow*, in "CVPR 2016 - IEEE Conference on Computer Vision & Pattern Recognition", LAS VEGAS, United States, June 2016, https://hal.archives-ouvertes.fr/hal-01240281

[11] V. KANTOROV, M. OQUAB, M. CHO, I. LAPTEV. *ContextLocNet: Context-Aware Deep Network Models for Weakly Supervised Localization*, in "ECCV 2016", Amsterdam, Netherlands, Springer, October 2016, pp. 350 - 365 [*DOI :* 10.1007/978-3-319-46454-1_22], https://hal.inria.fr/hal-01421772

[12] S. KWAK, M. CHO, I. LAPTEV. *Thin-Slicing for Pose: Learning to Understand Pose without Explicit Pose Estimation*, in "CVPR 2016 - IEEE Conference on Computer Vision and Pattern Recognition", Las Vegas, United States, June 2016, https://hal.inria.fr/hal-01242724

[13] 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

[14] G. SEGUIN, P. BOJANOWSKI, R. LAJUGIE, I. LAPTEV. *Instance-level video segmentation from object tracks*, in "CVPR 2016", Las Vegas, United States, Proceedings of the 29th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, June 2016, https://hal.inria.fr/hal-01255765

[15] G. A. SIGURDSSON, G. VAROL, X. WANG, A. FARHADI, I. LAPTEV, A. GUPTA. *Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding*, in "Computer Vision – ECCV 2016", Amsterdam, Netherlands, October 2016, pp. 510 - 526 [*DOI :* 10.1007/978-3-319-46448-0_31], https://hal.inria.fr/hal-01418216

[16] M. TRAGER, M. HEBERT, J. PONCE. *Consistency of silhouettes and their duals*, in "IEEE Conference on Computer Vision and Pattern Recognition, 2016", Las Vegas, United States, June 2016, https://hal.archives-ouvertes.fr/hal-01287180

[17] Y. ZHONG, R. ARANDJELOVIĆ, A. ZISSERMAN. *Faces In Places: Compound query retrieval*, in "BMVC - 27th British Machine Vision Conference", York, United Kingdom, September 2016, https://hal.inria.fr/hal-01353886

Other Publications

[18] A. BABENKO, R. ARANDJELOVIĆ, V. LEMPITSKY. *Pairwise Quantization*, June 2016, working paper or preprint, https://hal.inria.fr/hal-01330582

[19] J. PONCE, B. STURMFELS, M. TRAGER. *Congruences and Concurrent Lines in Multi-View Geometry*, 2017, Accepted for "Advances in Applied Mathematics", https://hal.inria.fr/hal-01423057

[20] G. VAROL, I. LAPTEV, C. SCHMID. *Long-term Temporal Convolutions for Action Recognition*, April 2016, working paper or preprint, https://hal.inria.fr/hal-01241518