IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2018

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:

A3.1.1. - Modeling, representation
A3.4. - Machine learning and statistics
A5.3. - Image processing and analysis
A5.4. - Computer vision
A9. - Artificial intelligence
A9.1. - Knowledge
A9.2. - Machine learning

Other Research Topics and Application Domains:

B9.5.1. - Computer science
B9.5.6. - Data science

1. Team, Visitors, External Collaborators

Research Scientists

Jean Ponce [Team leader, Inria, Senior Researcher, on leave from Ecole Normale Supérieure]
Ivan Laptev [Inria, Senior Researcher, HDR]
Josef Sivic [Inria, Senior Researcher, HDR]

Post-Doctoral Fellows

Justin Carpentier [Inria, from Sep 2018]
Vijay Kumar Reddy [Inria, from Jul 2018]
Sergey Zagoruyko [Inria, from Mar 2018]

PhD Students

Jean-Baptiste Alayrac [Inria, until Aug 2018]
Guilhem Cheron [Inria, until Aug 2018]
Theophile Dalens [Inria]
Thomas Eboli [Ecole Normale Superieure Paris]
Yana Hasson [Inria]
Yann Labbe [Ecole Normale Superieure Cachan, Intern from Apr 2018 then Phd from Sep 2018]
Zongmian Li [Inria]
Antoine Miech [Inria]
Julia Peyre [Inria]
Ronan Riochet [Inria]
Ignacio Rocco Spremolla [Inria]
Robin Strudel [Ecole Normale Superieure Paris, Intern from Apr 2018 then Phd from Oct 2018]
Matthew Trager [Inria, until Jun 2018]
Gul Varol [Inria]
Tuan Hung Vu [Inria, until Mar 2018]
Van Huy Vo [Ecole Normale Superieure Paris, from Dec 2018]
Dimitri Zhukov [Inria]

Technical staff

Sofiane Allayen [Inria, from May 2018]

Mauricio Diaz Melo [Inria, until Mar 2018]
Igor Kalevatykh [Inria]

Intern

Mihai Alexandru Dusmanu [Ecole Normale Superieure Paris, from Apr 2018]

Administrative Assistants

Helene Bessin Rousseau [Inria, from Mar 2018]
Sabrine Boumizy [Inria, until Feb 2018]
Helene Milome [Inria, from Dec 2018]

Visiting Scientists

Alexei Efros [UC Berkeley, from May 2018 until Jun 2018]
Ramazan Cinbis [Middle East Technical University, from Jul 2018 until Aug 2018]
David Fouhey [University of Michigan, from Sep 2018 until Nov 2018]
Pierre-Yves Masse [Czech Technical University, from Apr 2018]
Akihiko Torii [Tokyo Institute of Technology, from Apr 2018]

External Collaborator

Mathieu Aubry [Ecole Nationale des Ponts et Chaussees]

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, Ecole Normale Supérieure (ENS) and Centre National de la Recherche Scientifique (CNRS).

This year we have hired three new Phd students: Yann Labbe (ENS Cachan), Robin Strudel (ENS Lyon) and Van Huy Vo. Alexei Efros (Professor, UC Berkeley, USA) visited Willow during May-June. Ramazan Cinbis (Middle East Technical University) and David Fouhey (University of Michigan) visited Willow in July-August and September-November, respectively. Akihiko Torii (Tokyo Institute of Technology) spent sabbatical at Willow from Apr to August 2018. Finally, Pierre-Yves Masse (post-doc, Czech Technical University) spent 50% of his time at Sierra (F. Bach) and Willow teams as a visiting post-doc within the framework of collaboration with the Intelligent Machine Perception project lead by J. Sivic at the Czech Technical University in Prague.

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.

[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 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. Prizes and Awards

Antoine Miech, winner of a 2018 Google Fellowship.

5.1.2. Visibility

- J. Ponce co-organized the PRAIRIE AI Summer School, Grenoble, 2018, which brought together 200 participants representing 44 different nationalities, and selected from 700 applications, with 60% students, 15% academics , and 25% industrials. 25% of these participants were women.
- I. Laptev served as Program Chair for the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018. CVPR is the largest computer vision conference. The 2018 edition has 3,309 paper submissions, 979 accepted papers and 6,128 registered attendees.
- J. Ponce has been a key person in creating the PRAIRIE Institute for AI research in Paris, announced on the occasion of the AI for Humanity summit organized by President Emmanuel Macron in 2018 (https://www.inria.fr/en/news/news-from-inria/launch-of-the-prairie-institute). He has also been a key player in bringing together its industrial and international partners.

6. New Software and Platforms

6.1. NCNet

Neighbourhood Consensus Networks

KEYWORDS: Computer vision - Machine learning

FUNCTIONAL DESCRIPTION: Open source release of the software package for the NIPS'18 paper by Rocco et al. "Neighbourhood Consensus Networks". This release provides a full implementation of the method, including code for training models, and testing on standard datasets, as well as trained models.

- Participants: Ignacio Rocco Spremolla, Mircea Cimpoi, Akihiko Torii, Relja Arandjelovic, Tomas Pajdla and Josef Sivic
- Contact: Ignacio Rocco Spremolla
- Publication: Neighbourhood Consensus Networks
- URL: https://www.di.ens.fr/willow/research/ncnet/

6.2. Mixture-of-Embedding-Experts

KEYWORD: Computer vision
FUNCTIONAL DESCRIPTION: Joint understanding of video and language is an active research area with many applications. Prior work in this domain typically relies on learning text-video embeddings. One difficulty with this approach, however, is the lack of large-scale annotated video-caption datasets for training. To address this issue, we aim at learning text-video embeddings from heterogeneous data sources. To this end, we propose a Mixture-of-Embedding-Experts (MEE) model with ability to handle missing input modalities during training. As a result, our framework can learn improved text-video embeddings simultaneously from image and video datasets. We also show the generalization of MEE to other input modalities such as face descriptors.

- Participants: Ivan Laptev and Josef Sivic
- Contact: Antoine Miech
- Publication: Learning a Text-Video Embedding from Incomplete and Heterogeneous Data
- URL: https://www.di.ens.fr/willow/research/mee/

6.3. BodyNet

BodyNet: Volumetric Inference of 3D Human Body Shapes
KEYWORDS: Computer vision - Machine learning
FUNCTIONAL DESCRIPTION: BodyNet has the code to train multi-task neural networks for 2D/3D pose estimation, 2D body part segmentation, and 3D volumetric shape estimation of human bodies given single RGB images as input. The release also contains pre-trained models.

- Participants: Gül Varol Simsekli, Ivan Laptev and Cordelia Schmid
- Contact: Gül Varol Simsekli
- Publication: BodyNet: Volumetric Inference of 3D Human Body Shapes
- URL: https://www.di.ens.fr/willow/research/bodynet/

6.4. FlexWLoc

Flexible Weakly supervised action Localization model
KEYWORDS: Computer vision - Machine learning
FUNCTIONAL DESCRIPTION: Open source release of the software package for the NIPS'18 paper by Chéron et al. "A flexible model for training action localization with varying levels of supervision". This release provides a full implementation of the method, including code for training and testing.

- Participants: Guilhem Chéron, Jean-Baptiste Alayrac, Ivan Laptev and Cordelia Schmid
- Contact: Guilhem Chéron
- Publication: A flexible model for training action localization with varying levels of supervision
- URL: https://www.di.ens.fr/willow/research/weakactionloc/

6.5. Pinocchio

KEYWORDS: Robotics - Biomechanics - Mechanical multi-body systems
FUNCTIONAL DESCRIPTION: Pinocchio instantiates state-of-the-art Rigid Body Algorithms for poly-articulated systems based on revisited Roy Featherstone's algorithms. In addition, Pinocchio instantiates analytical derivatives of the main Rigid-Body Algorithms like the Recursive Newton-Euler Algorithms or the Articulated-Body Algorithm. Pinocchio is first tailored for legged robotics applications, but it can be used in extra contexts. It is built upon Eigen for linear algebra and FCL for collision detection. Pinocchio comes with a Python interface for fast code prototyping.

- Partner: CNRS
- Contact: JUSTIN CARPENTIER
- URL: https://github.com/stack-of-tasks/pinocchio

6.6. weakalign

End-to-end weakly-supervised semantic alignment
KEYWORDS: Computer vision - Machine learning
FUNCTIONAL DESCRIPTION: Open source release of the software package for the CVPR'18 paper by Rocco et al. "End-to-end weakly-supervised semantic alignment". This release provides a full implementation of the method, including code for training models, and testing on standard datasets, as well as trained models.

- Participants: Ignacio Rocco Spremolla, Relja Arandjelovic and Josef Sivic
- Contact: Ignacio Rocco Spremolla
- Publication: End-to-end weakly-supervised semantic alignment
- URL: https://www.di.ens.fr/willow/research/weakalign/

6.7. InLoc

Indoor Visual Localization with Dense Matching and View Synthesis
KEYWORD: Computer vision
FUNCTIONAL DESCRIPTION: Open source release of the software package for the CVPR'18 paper by Taira et al. "Indoor Visual Localization with Dense Matching and View Synthesis". This release provides a full implementation of the method.

- Participants: Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla and Akihiko Torii
- Contact: Josef Sivic
- Publication: InLoc: Indoor Visual Localization with Dense Matching and View Synthesis
- URL: https://github.com/HajimeTaira/InLoc_demo

7. New Results

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

7.1.1. Indoor Visual Localization with Dense Matching and View Synthesis

Participants: Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, Akihiko Torii.

In [20], we seek to predict the 6 degree-of-freedom (6DoF) pose of a query photograph with respect to a large indoor 3D map. The contributions of this work are three-fold. First, we develop a new large-scale visual localization method targeted for indoor environments. The method proceeds along three steps: (i) efficient retrieval of candidate poses that ensures scalability to large-scale environments, (ii) pose estimation using dense matching rather than local features to deal with textureless indoor scenes, and (iii) pose verification by virtual view synthesis to cope with significant changes in viewpoint, scene layout, and occluders. Second, we collect a new dataset with reference 6DoF poses for large-scale indoor localization. Query photographs are captured by mobile phones at a different time than the reference 3D map, thus presenting a realistic indoor localization scenario. Third, we demonstrate that our method significantly outperforms current state-of-the-art indoor localization approaches on this new challenging data. Figure 1 presents some example results.

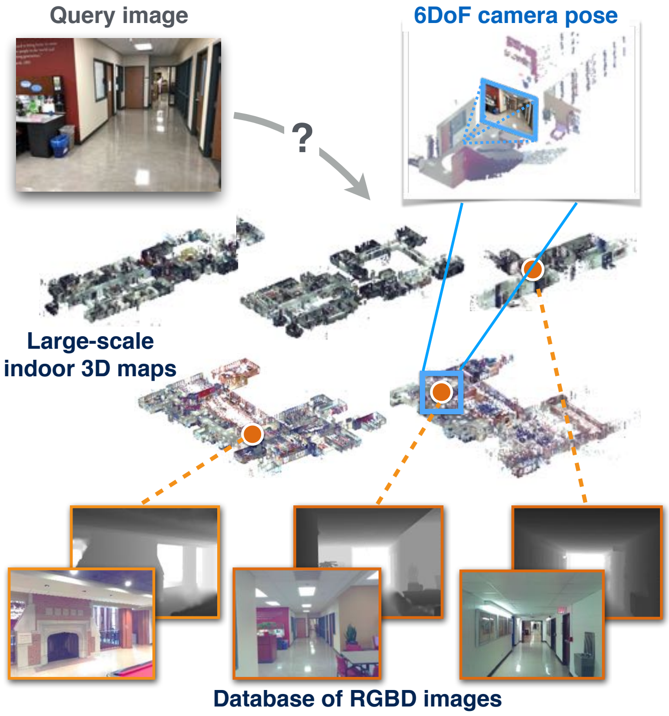

Figure 1. Large-scale indoor visual localization. Given a database of geometrically-registered RGBD images, we predict the 6DoF camera pose of a query RGB image by retrieving candidate images, estimating candicate camera poses, and selecting the best matching camera pose. To address inherent difficulties in indoor visual localization, we introduce the ?InLoc? approach that performs a sequence of progressively stricter verification steps.

7.1.2. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions

Participants: Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Frederik Kahl, Tomas Pajdla.

Visual localization enables autonomous vehicles to navigate in their surroundings and augmented reality applications to link virtual to real worlds. Practical visual localization approaches need to be robust to a wide variety of viewing condition, including day-night changes, as well as weather and seasonal variations, while providing highly accurate 6 degree-of-freedom (6DOF) camera pose estimates. In [19], we introduce the first benchmark datasets specifically designed for analyzing the impact of such factors on visual localization. Using carefully created ground truth poses for query images taken under a wide variety of conditions, we evaluate the impact of various factors on 6DOF camera pose estimation accuracy through extensive experiments with state-of-the-art localization approaches. Based on our results, we draw conclusions about the difficulty of different conditions, showing that long-term localization is far from solved, and propose promising avenues for future work, including sequence-based localization approaches and the need for better local features. Our benchmark is available at visuallocalization.net. Figure 2 presents some example results.

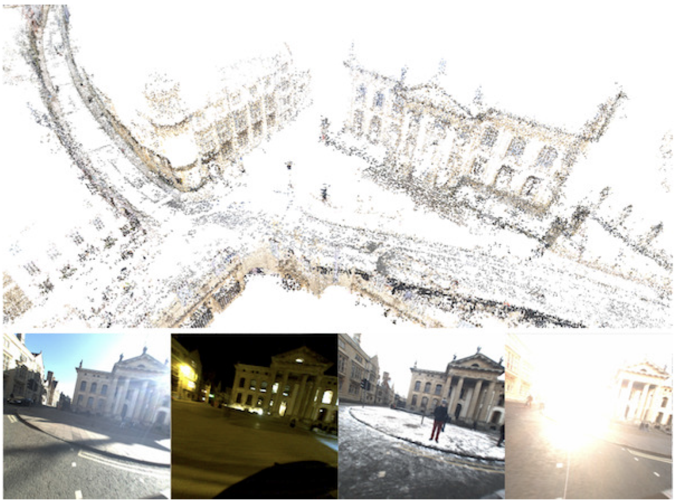

Figure 2. Visual localization in changing urban conditions. We present three new datasets, Aachen Day-Night, RobotCar Seasons (shown) and CMU Seasons for evaluating 6DOF localization against a prior 3D map (top) using registered query images taken from a wide variety of conditions (bottom), including day-night variation, weather, and seasonal changes over long periods of time.

7.1.3. Changing Views on Curves and Surfaces

Participants: Kathlen Kohn, Bernd Sturmfels, Matthew Trager, Boris Bukh, Xavier Goaoc, Alfredo Hubard, Matthew Trager.

Visual events in computer vision are studied from the perspective of algebraic geometry. Given a sufficiently general curve or surface in 3-space, we consider the image or contour curve that arises by projecting from a viewpoint. Qualitative changes in that curve occur when the viewpoint crosses the visual event surface as illustrated in 3. We examine the components of this ruled surface, and observe that these coincide with the iterated singular loci of the coisotropic hypersurfaces associated with the original curve or surface. We derive formulas, due to Salmon and Petitjean, for the degrees of these surfaces, and show how to compute exact representations for all visual event surfaces using algebraic methods. This work has been published in [8].

subsectionConsistent Sets of Lines with no Colorful Incidence

We consider incidences among colored sets of lines in $\mathbb{R}^d$ and examine whether the existence of certain concurrences between lines of k colors force the existence of at least one concurrence between lines of $k + 1$ colors. This question is relevant for problems in 3D reconstruction in computer vision such as the one illustrated in Figure 4. This work has been published in [12].

7.1.4. On the Solvability of Viewing Graphs

Participants: Matthew Trager, Brian Osserman, Jean Ponce.

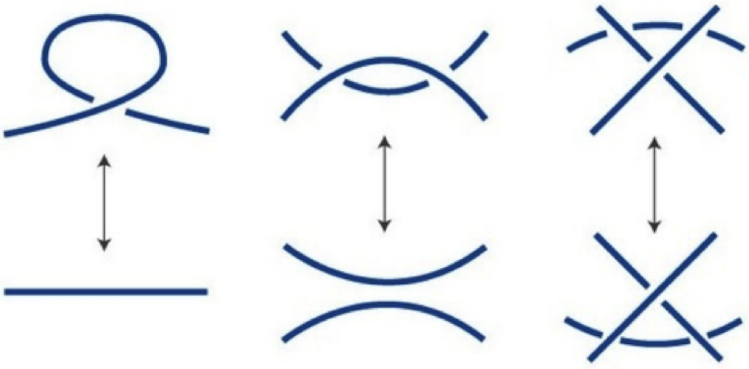

Figure 3. Changing views of a curve correspond to Reidemeister moves. The viewpoint z crosses the tangential surface (left), edge surface (middle), or trisecant surface (right).

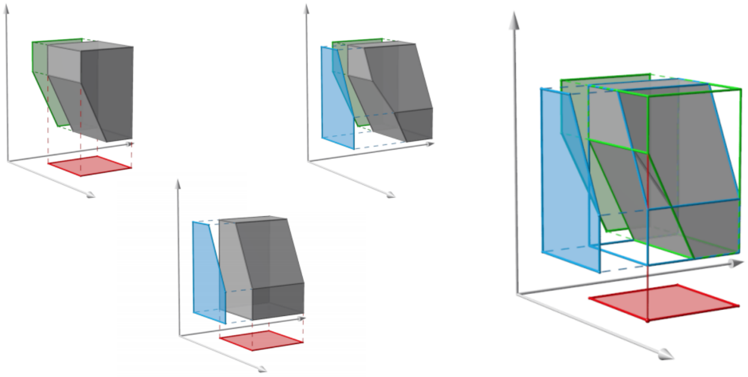

Figure 4. Three silhouettes that are 2-consistent but not globally consistent for three orthogonal projections. Each of the first three figures shows a three-dimensional set that projects onto two of the three silhouettes. The fourth figure illustrates that no set can project simultaneously onto all three silhouettes: the highlighted red image point cannot be lifted in 3D, since no point that projects onto it belongs to the pre-images of both the blue and green silhouettes.

A set of fundamental matrices relating pairs of cameras in some configuration can be represented as edges of a ” viewing graph ". Whether or not these fundamental matrices are generically sufficient to recover the global camera configuration depends on the structure of this graph. We study characterizations of ” solvable " viewing graphs, and present several new results that can be applied to determine which pairs of views may be used to recover all camera parameters. We also discuss strategies for verifying the solvability of a graph computationally. This work has been published in [21].

7.1.5. *In Defense of Relative Multi-View Geometry*

Participants: Matthew Trager, Jean Ponce.

The idea of studying multi-view geometry and structure-from-motion problems *relative* to the scene and camera configurations, without appeal to external coordinate systems, dates back to the early days of modern geometric computer vision. Yet, it has a bad rap, the scene reconstructions obtained often being deemed as inaccurate despite careful implementations. The aim of this article is to correct this perception with a series of new results. In particular, we show that using a small subset of scene and image points to parameterize their relative configurations offers a natural coordinate-free formulation of Carlsson-Weinshall duality for arbitrary numbers of images. An example is shown in Figure 5. For three views, this approach also yields novel purely- and quasi-linear formulations of structure from motion using *reduced trilinearities*, without the complex polynomial constraints associated with trifocal tensors, revealing in passing the strong link between “3D” ($\mathbb{P}^3 \to \mathbb{P}^2$) and “2D” ($\mathbb{P}^2 \to \mathbb{P}^1$) models of trinocular vision. Finally, we demonstrate through preliminary experiments that the proposed relative reconstruction methods gives good results on real data. This works is available as a preprint [32].

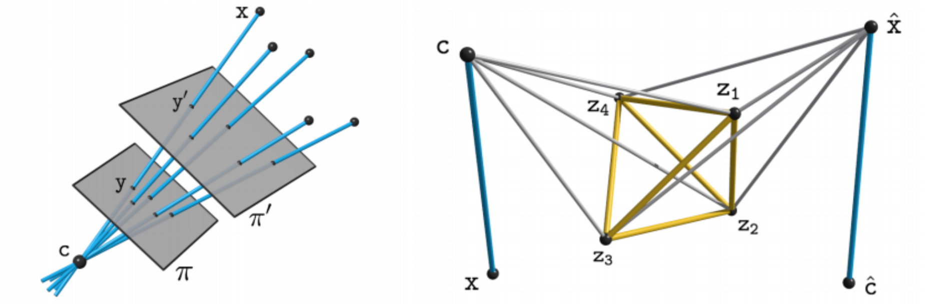

Figure 5. Configurations. ***Left****: Image point and viewing ray configurations are isomorphic and independent of the retinal plane.* ***Right****: Geometric Carlsson-Weinshall duality between scene point and pinhole configurations.*

7.1.6. *Multigraded Cayley-Chow Forms*

Participants: Brian Osserman, Matthew Trager.

We introduce a theory of multigraded Cayley-Chow forms associated to subvarieties of products of projective spaces. Figure 6 illustrares some examples of projective speces. Two new phenomena arise: first, the construction turns out to require certain inequalities on the dimensions of projections; and second, in positive characteristic the multigraded Cayley-Chow forms can have higher multiplicities. The theory also provides a natural framework for understanding multifocal tensors in computer vision. This works is available as a preprint [30].

7.2. Category-level object and scene recognition

7.2.1. *Detecting rare visual relations using analogies*

Participants: Julia Peyre, Cordelia Schmid, Ivan Laptev, Josef Sivic.

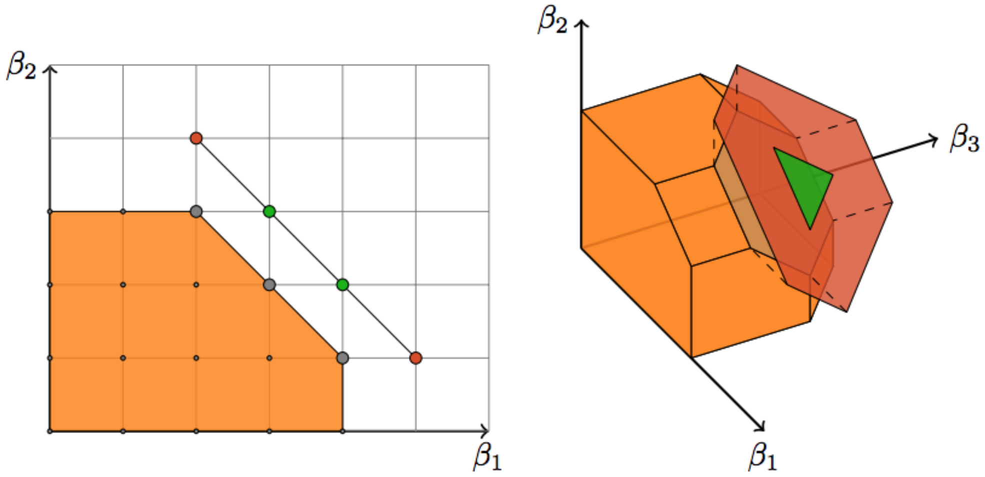

Figure 6. Two polymatroids. The sets of bases (corresponding to our multidegree supports) are in gray; while the sets of circuits and of non-circuit 1-deficient vectors are in green and red, respectively.

We seek to detect visual relations in images of the form of triplets t = (subject, predicate, object), such as "person riding dog", where training examples of the individual entities are available but their combinations are rare or unseen at training such as shown in Figure 7. This is an important set-up due to the combinatorial nature of visual relations : collecting sufficient training data for all possible triplets would be very hard. The contributions of this work are three-fold. First, we learn a representation of visual relations that combines (i) individual embeddings for subject, object and predicate together with (ii) a visual phrase embedding that represents the relation triplet. Second, we learn how to transfer visual phrase embeddings from existing training triplets to unseen test triplets using analogies between relations that involve similar objects. Third, we demonstrate the benefits of our approach on two challenging datasets involving rare and unseen relations : on HICO-DET, our model achieves significant improvement over a strong baseline, and we confirm this improvement on retrieval of unseen triplets on the UnRel rare relation dataset. This work, currently under review, can be found at [31].

7.2.2. *Convolutional neural network architecture for geometric matching*

Participants: Ignacio Rocco, Relja Arandjelović, Josef Sivic.

In [9], we address the problem of determining correspondences between two images in agreement with a geometric model such as an affine, homography or thin-plate spline transformation, and estimating its parameters. The contributions of this work are threefold. First, we propose a convolutional neural network architecture for geometric matching. The architecture is based on three main components that mimic the standard steps of feature extraction, matching and simultaneous inlier detection and model parameter estimation, while being trainable end-to-end. Second, we demonstrate that the network parameters can be trained from synthetically generated imagery without the need for manual annotation and that our matching layer significantly increases generalization capabilities to never seen before images. Finally, we show that the same model can perform both instance-level and category-level matching giving state-of-the-art results on the challenging PF, TSS and Caltech-101 datasets.

7.2.3. *End-to-end weakly-supervised semantic alignment*

Participants: Ignacio Rocco, Relja Arandjelović, Josef Sivic.

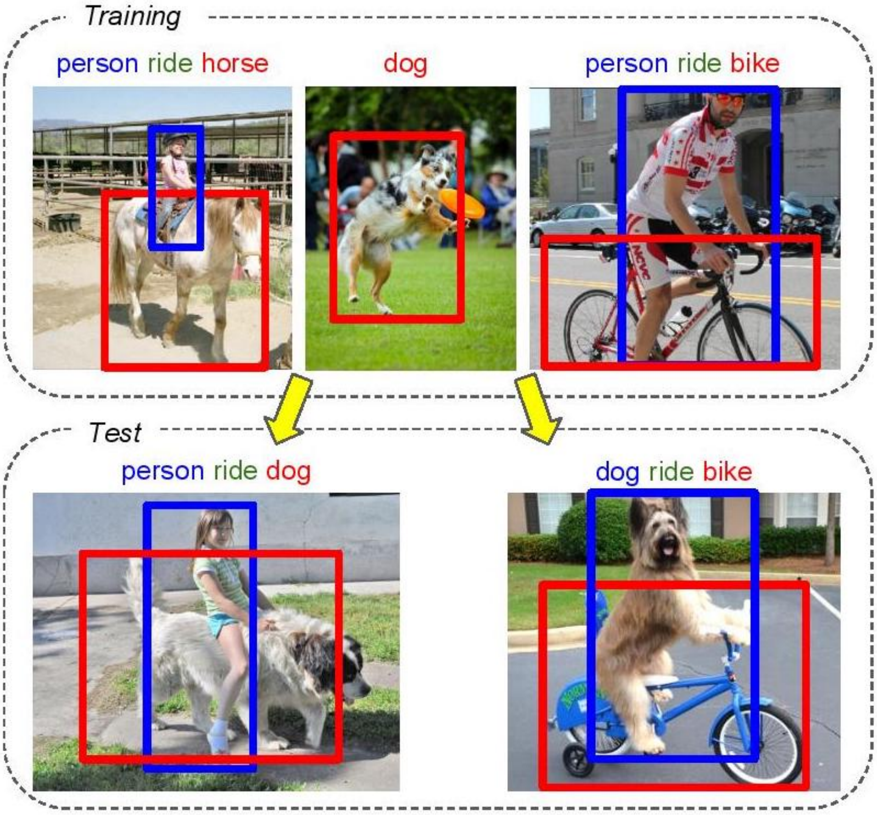

Figure 7. Illustration of transfer by analogy from seen training triplets (e.g. "person ride horse") to unseen or rare ones (e.g. "person ride dog"))

In [17], we tackle the task of semantic alignment where the goal is to compute dense semantic correspondence aligning two images depicting objects of the same category. This is a challenging task due to large intra-class variation, changes in viewpoint and background clutter. We present the following three principal contributions. First, we develop a convolutional neural network architecture for semantic alignment that is trainable in an end-to-end manner from weak image-level supervision in the form of matching image pairs. The outcome is that parameters are learnt from rich appearance variation present in different but semantically related images without the need for tedious manual annotation of correspondences at training time. Second, the main component of this architecture is a differentiable soft inlier scoring module, inspired by the RANSAC inlier scoring procedure, that computes the quality of the alignment based on only geometrically consistent correspondences thereby reducing the effect of background clutter. Third, we demonstrate that the proposed approach achieves state-of-the-art performance on multiple standard benchmarks for semantic alignment. Figure 8 presents some example results.

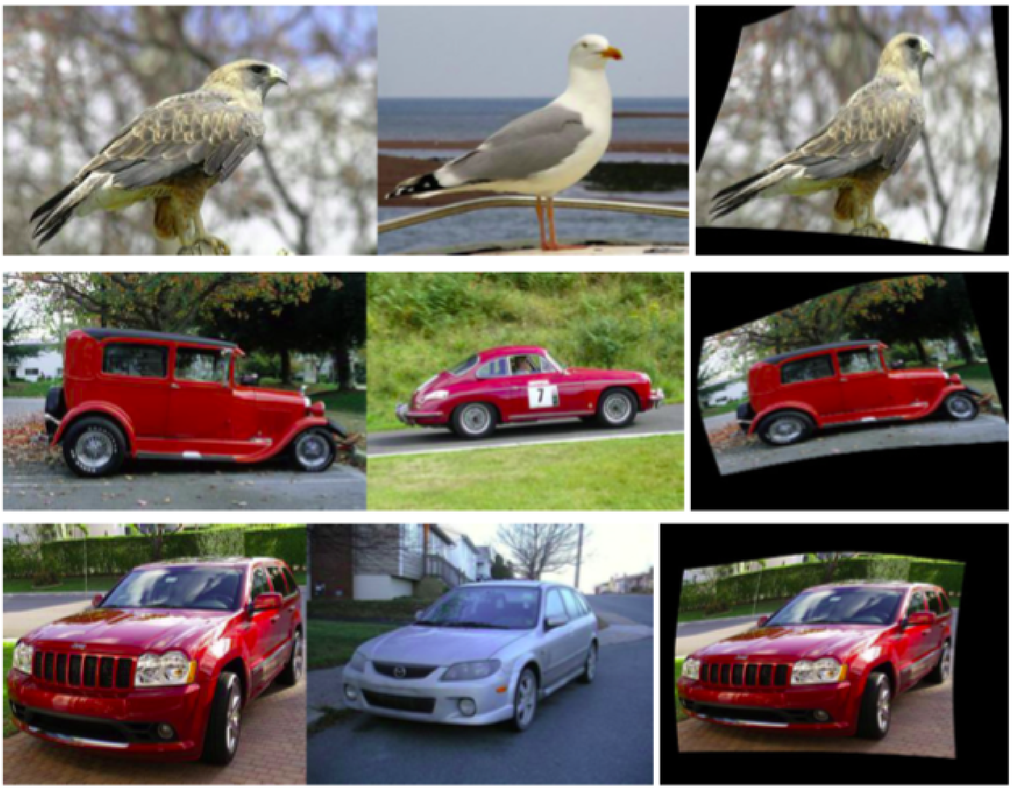

Figure 8. Each row corresponds to one example and shows the (right) automatic semantic alignment of the (left) source and (middle) target images.

7.2.4. Neighbourhood Consensus Networks

Participants: Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, Josef Sivic.

In [18], we address the problem of finding reliable dense correspondences between a pair of images. This is a challenging task due to strong appearance differences between the corresponding scene elements and ambiguities generated by repetitive patterns. The contributions of this work are threefold. First, inspired by the classic idea of disambiguating feature matches using semi-local constraints, we develop an end-to-end trainable convolutional neural network architecture that identifies sets of spatially consistent matches by analyzing neighbourhood consensus patterns in the 4D space of all possible correspondences between a pair of images without the need for a global geometric model. Second, we demonstrate that the model can be trained effectively from weak supervision in the form of matching and non-matching image pairs without the need for costly manual annotation of point to point correspondences. Third, we show the proposed neighbourhood consensus network can be applied to a range of matching tasks including both category- and instance-level matching, obtaining the state-of-the-art results on the PF Pascal dataset and the InLoc indoor visual localization benchmark. Figure 9 shows the network architecture of the proposed Neighbourhood Consensus Network, that features 3 layers of 4D convolutions.

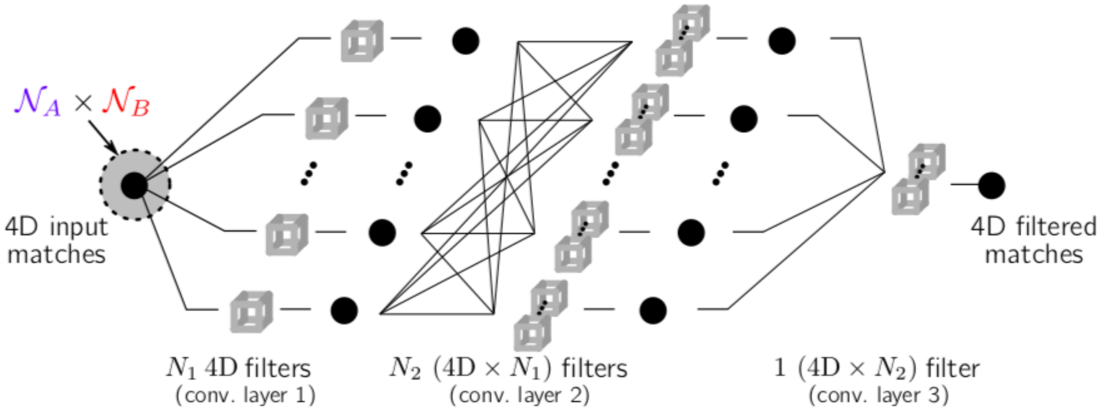

Figure 9. A neighbourhood consensus CNN operates on the 4D space of feature matches. The first 4D convolutional layer filters span $\mathcal{N}_A \times \mathcal{N}_B$, the Cartesian product of local neighbourhoods $\mathcal{N}_A$ and $\mathcal{N}_B$ in images A and B respectively. The proposed 4D neighbourhood consensus CNN can learn to identify the matching patterns of reliable and unreliable matches, and filter the matches accordingly

7.2.5. Compressing the Input for CNNs with the First-Order Scattering Transform

Participants: Edouard Oyallon, Eugene Belilovsky, Sergey Zagoruyko, Michal Valko.

In [16], we study the first-order scattering transform as a candidate for reducing the signal processed by a convolutional neural network (CNN). We study this transformation and show theoretical and empirical evidence that in the case of natural images and sufficiently small translation invariance, this transform preserves most of the signal information needed for classification while substantially reducing the spatial resolution and total signal size. We show that cascading a CNN with this representation performs on par with ImageNet classification models commonly used in downstream tasks such as the ResNet-50. We subsequently apply our trained hybrid ImageNet model as a base model on a detection system, which has typically larger image inputs. On Pascal VOC and COCO detection tasks we deliver substantial improvements in the inference speed and training memory consumption compared to models trained directly on the input image.

7.2.6. Exploring Weight Symmetry in Deep Neural Networks

Participants: Xu Shell Hu, Sergey Zagoruyko, Nikos Komodakis.

In [27], we propose to impose symmetry in neural network parameters to improve parameter usage and make use of dedicated convolution and matrix multiplication routines. Due to significant reduction in the number of parameters as a result of the symmetry constraints, one would expect a dramatic drop in accuracy. Surprisingly, we show that this is not the case, and, depending on network size, symmetry can have little or no negative effect on network accuracy, especially in deep overparameterized networks. We propose several ways to impose local symmetry in recurrent and convolutional neural networks, and show that our symmetry parameterizations satisfy universal approximation property for single hidden layer networks. We extensively evaluate these parameterizations on CIFAR, ImageNet and language modeling datasets, showing significant benefits from the use of symmetry. For instance, our ResNet-101 with channel-wise symmetry has almost 25% less parameters and only 0.2% accuracy loss on ImageNet.

7.3. Image restoration, manipulation and enhancement

7.3.1. Neural Embedding of an Iterative Deconvolution Algorithm for Motion Blur Estimation and Removal

Participants: Thomas Eboli, Jian Sun, Jean Ponce.

We introduce a new two-steps learning-based approach to motion blur estimation and removal decomposed into two trainable modules. A local linear motion model is estimated at each pixel using a first convolutional neural network (CNN) in a regression setting. It is then used to drive an algorithm that casts non-blind, non-uniform image deblurring as a least-squares problem regularized by natural image priors in the form of sparsity constraints. This problem is solved by combining the alternative direction method of multipliers with an iterative residual compensation algorithm, with a finite number of iterations embedded into a second CNN whose trainable parameters are deconvolution filters. The second network outputs the sharp image, and the two CNNs can be trained together in an end-to-end manner. Our experiments demonstrate that the proposed method is significantly faster than existing ones, and provides competitive results with the state of the art on synthetic and real data. This work is available as a pre-print[25] and an example is illustrated in Figure 10.

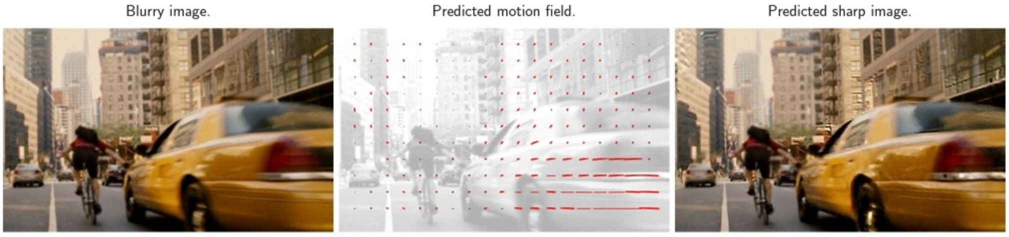

Figure 10. From a blurry image, we first use CNN-based regressor to predict a motion field with local linear motions before using it in a trainable iterative residual compensation algorithm to restore the image.

7.3.2. Deformable Kernel Networks for Joint Image Filtering

Participants: Beomjun Kim, Jean Ponce, Bumsub Ham.

Joint image filters are used to transfer structural details from a guidance picture used as a prior to a target image, in tasks such as enhancing spatial resolution and suppressing noise. Previous methods based on convolutional neural networks (CNNs) combine nonlinear activations of spatially-invariant kernels to estimate structural details and regress the filtering result. In this paper, we instead learn explicitly sparse and spatially-variant kernels. We propose a CNN architecture and its efficient implementation, called the deformable kernel network (DKN), that outputs sets of neighbors and the corresponding weights adaptively for each pixel. The filtering result is then computed as a weighted average. We also propose a fast version of DKN that runs about four times faster for an image of size 640 by 480. We demonstrate the effectiveness and flexibility of our models on the tasks of depth map upsampling, saliency map upsampling, cross-modality image restoration, texture removal, and semantic segmentation. In particular, we show that the weighted averaging process with sparsely sampled 3 by 3 kernels outperforms the state of the art by a significant margin. This works has been submitted to the IEEE Trans. on Pattern Analysis and Machine Intelligence and is available as a pre-print [28].

7.4. Human activity capture and classification

7.4.1. Learning a Text-Video Embedding from Incomplete and Heterogeneous Data

Participants: Antoine Miech, Ivan Laptev, Josef Sivic.

Joint understanding of video and language is an active research area with many applications. Prior work in this domain typically relies on learning text-video embeddings. One difficulty with this approach, however, is the lack of large-scale annotated video-caption datasets for training. To address this issue, in [29] we aim at learning text-video embeddings from heterogeneous data sources. To this end, we propose a Mixture-of-Embedding-Experts (MEE) model with ability to handle missing input modalities during training. As a result, our framework can learn improved text-video embeddings simultaneously from image and video datasets. We also show the generalization of MEE to other input modalities such as face descriptors. We evaluate our method on the task of video retrieval and report results for the MPII Movie Description and MSR-VTT datasets. The proposed MEE model demonstrates significant improvements and outperforms previously reported methods on both text-to-video and video-to-text retrieval tasks. Figure 11 illustrates application of our method in text-to-video retrieval.

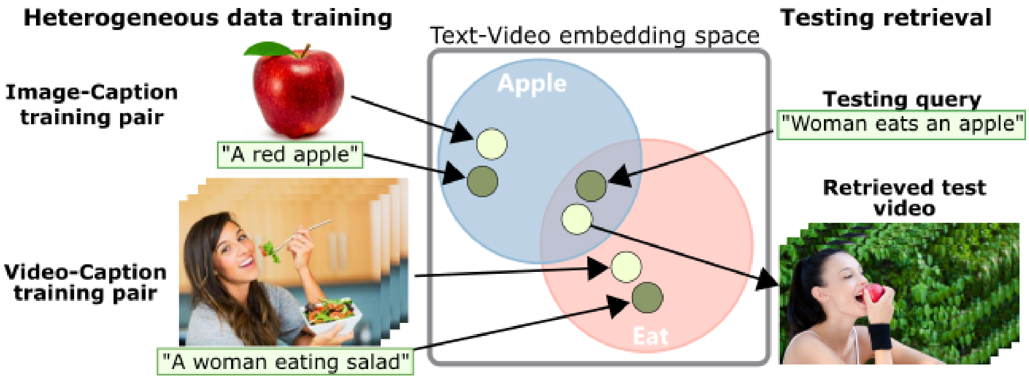

Figure 11. We learn a text-video embedding from heterogenous (here Image-Text and Video-Text) data sources. At test time, we can query concepts learnt from both Image-Caption and Video-Caption training pair (e.g. the eating notion being learnt from video and the apple notion from image).

7.4.2. A flexible model for training action localization with varying levels of supervision

Participants: Guilhem Chéron, Jean-Baptiste Alayrac, Ivan Laptev, Cordelia Schmid.

Spatio-temporal action detection in videos is typically addressed in a fully-supervised setup with manual annotation of training videos required at every frame. Since such annotation is extremely tedious and prohibits scalability, there is a clear need to minimize the amount of manual supervision. In this work we propose a unifying framework that can handle and combine varying types of less-demanding weak supervision. Our model is based on discriminative clustering and integrates different types of supervision as constraints on the optimization as illustrated in Figure 12. We investigate applications of such a model to training setups with alternative supervisory signals ranging from video-level class labels to the full per-frame annotation of action bounding boxes. Experiments on the challenging UCF101-24 and DALY datasets demonstrate competitive performance of our method at a fraction of supervision used by previous methods. The flexibility of our model enables joint learning from data with different levels of annotation. Experimental results demonstrate a significant gain by adding a few fully supervised examples to otherwise weakly labeled videos. This work has been published in [14].

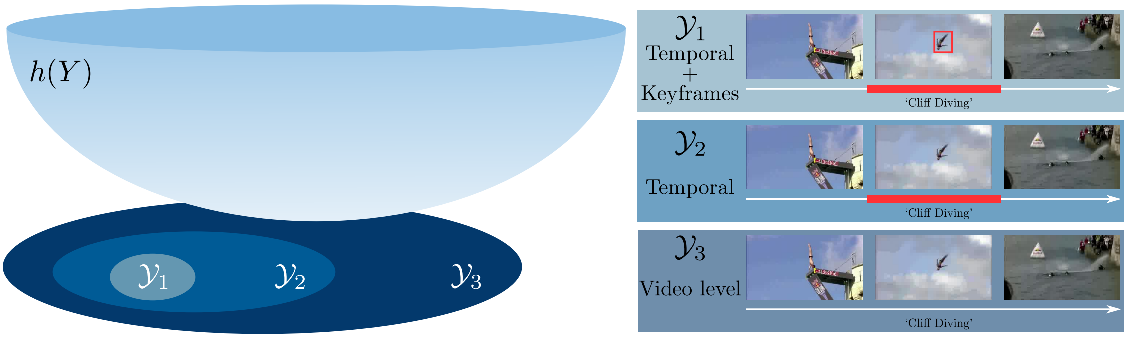

Figure 12. Our method estimates a matrix Y assigning human tracklets to action labels in training videos by optimizing an objective function $h(Y)$ under constraints $\mathcal{Y}_s$. Different types of supervision define particular constraints $\mathcal{Y}_s$ and do not affect the form of the objective function. The increasing level of supervision imposes stricter constraints, e.g. $\mathcal{Y}_1 \supset \mathcal{Y}_2 \supset \mathcal{Y}_3 \supset \mathcal{Y}_4$ as illustrated for the Cliff Diving example above.

7.4.3. BodyNet: Volumetric Inference of 3D Human Body Shapes

Participants: Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, Cordelia Schmid.

Human shape estimation is an important task for video editing, animation and fashion industry. Predicting 3D human body shape from natural images, however, is highly challenging due to factors such as variation in human bodies, clothing and viewpoint. Prior methods addressing this problem typically attempt to fit parametric body models with certain priors on pose and shape. In this work we argue for an alternative representation and propose BodyNet, a neural network for direct inference of volumetric body shape from a single image. BodyNet is an end-to-end trainable network that benefits from (i) a volumetric 3D loss, (ii) a multi-view re-projection loss, and (iii) intermediate supervision of 2D pose, 2D body part segmentation, and 3D pose. Each of them results in performance improvement as demonstrated by our experiments. To evaluate the method, we fit the SMPL model to our network output and show state-of-the-art results on the SURREAL and Unite the People datasets, outperforming recent approaches. Besides achieving state-of-the-art performance, our method also enables volumetric body-part segmentation. Figure 13 illustrates the volumetric outputs given two sample input images. This work has been published at ECCV 2018 [22].

7.4.4. Localizing Moments in Video with Temporal Language

Participants: Lisa Anne Hendricks, Oliver Wang, Eli Schechtman, Josef Sivic, Trevor Darrell, Bryan Russell.

Localizing moments in a longer video via natural language queries is a new, challenging task at the intersection of language and video understanding. Though moment localization with natural language is similar to other language and vision tasks like natural language object retrieval in images, moment localization offers an interesting opportunity to model temporal dependencies and reasoning in text. In [15], we propose a new model that explicitly reasons about different temporal segments in a video, and shows that temporal context

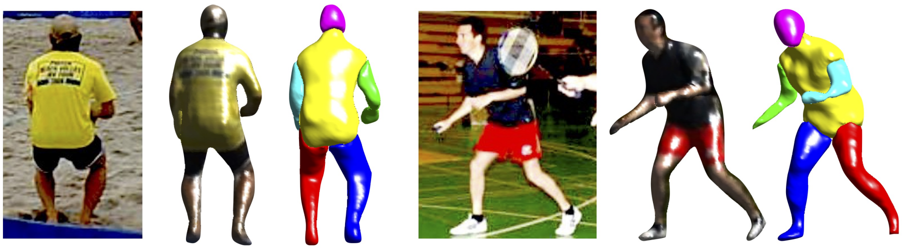

Figure 13. Our BodyNet predicts a volumetric 3D human body shape and 3D body parts from a single image. We show the input image, the predicted human voxels, and the predicted part voxels.

is important for localizing phrases which include temporal language. To benchmark whether our model, and other recent video localization models, can effectively reason about temporal language, we collect the novel TEMPO-ral reasoning in video and language (TEMPO) dataset. Our dataset consists of two parts: a dataset with real videos and template sentences (TEMPO - Template Language) which allows for controlled studies on temporal language, and a human language dataset which consists of temporal sentences annotated by humans (TEMPO - Human Language).

7.4.5. The Pinocchio C++ library ? A fast and flexible implementation of rigid body dynamics algorithms and their analytical derivatives

Participants: Justin Carpentier, Guilhem Saurel, Gabriele Buondonno, Joseph Mirabel, Florent Lamiraux, Olivier Stasse, Nicolas Mansard.

In this work, we introduce Pinocchio, an open-source software framework that implements rigid body dynamics algorithms and their analytical derivatives. Pinocchio does not only include standard algorithms employed in robotics (e.g., forward and inverse dynamics) but provides additional features essential for the control, the planning and the simulation of robots. In this paper, we describe these features and detail the programming patterns and design which make Pinocchio efficient. We evaluate the performances against RBDL, another framework with broad dissemination inside the robotics community. We also demonstrate how the source code generation embedded in Pinocchio outperforms other approaches of state of the art.

7.4.6. Modeling Spatio-Temporal Human Track Structure for Action Localization

Participants: Guilhem Chéron, Anton Osokin, Ivan Laptev, Cordelia Schmid.

This paper [24] addresses spatio-temporal localization of human actions in video. In order to localize actions in time, we propose a recurrent localization network (RecLNet) designed to model the temporal structure of actions on the level of person tracks. Our model is trained to simultaneously recognize and localize action classes in time and is based on two layer gated recurrent units (GRU) applied separately to two streams, i.e. appearance and optical flow streams. When used together with state-of-the-art person detection and tracking, our model is shown to improve substantially spatio-temporal action localization in videos. The gain is shown to be mainly due to improved temporal localization as illustrated in Figure 14. We evaluate our method on two recent datasets for spatio-temporal action localization, UCF101-24 and DALY, demonstrating a significant improvement of the state of the art.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

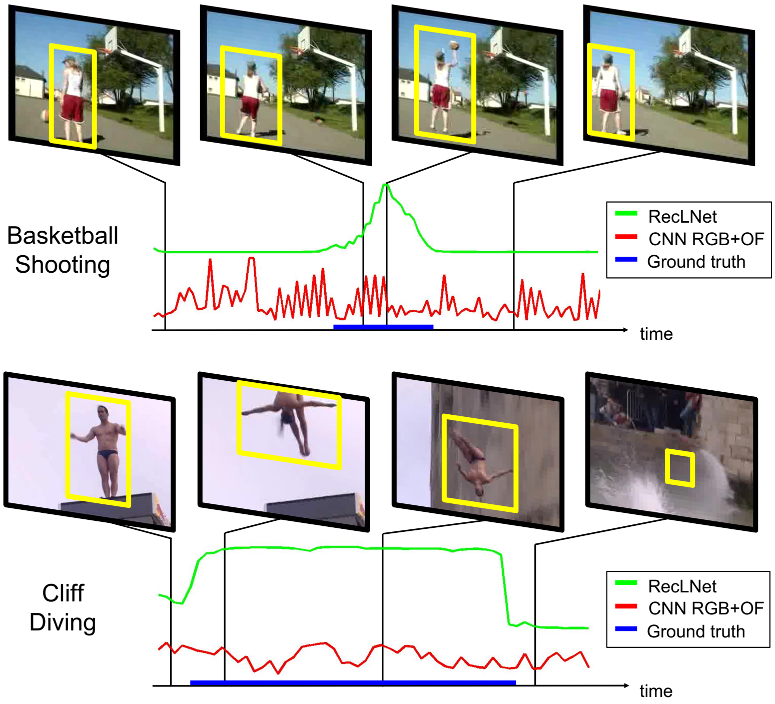

Figure 14. Spatio-temporal action localization using a CNN baseline (red) and our RecLNet (green) both applied on the level of person tracks. Our approach provides accurate temporal boundaries when the action happens.

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

Participants: Guilhem Cheron, 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 Sciencea 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 2013-2017 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.

In 2018 a new agreement has been signed with a new focus on video understanding for personal assistants. The scientific objectives are to develop models, representations and learning algorithms for (i) automatic understanding of task-driven complex human activities from videos narrated with natural language in order to (ii) give people instructions in a new environment via an augmented reality device such as the Microsoft HoloLens. Besides the clear scientific interest of automatically understanding human activities in video streams, the main high-impact motivation of this project it to develop virtual assistants that may guide a child through simple games to improve his/her manipulation and language skills; help an elderly person to achieve everyday tasks; or facilitate the training of a new worker for highly-specialized machinery maintenance.

8.1.2. Louis Vuitton/ENS chair on artificial intelligence

Participants: Ivan Laptev, Jean Ponce, Josef Sivic.

The scientific chair Louis Vuitton - École normale supérieure in Artificial Intelligence has been created in 2017 and inaugurated on April 12, 2018 by the ENS Director Marc Mézard and the LV CEO Michael Burke. The goal of the chair is to establish a close collaboration between LV and ENS in the area of Artificial Intelligence. The chair enjoys the generous annual contribution of 200K Euros provided by LV in support of research activities in statistical learning and computer vision. In particular, the chair supports the costs of researchers, students, missions, computational resources as well as seminars and meetings, including the two days of meeting annually organized by LV and ENS. During 2018 ENS and LV have organized several joint meetings with the participation of researchers from SIERRA and WILLOW teams. The chair has also supported the hiring of one PhD student at the WILLOW team, missions to conferences and international research labs as well as data collection for research projects.

8.2. Bilateral Grants with Industry

8.2.1. 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.

8.2.2. 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.

8.2.3. 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.

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. DGA - RAPID project DRAAF

Participant: Ivan Laptev.

DGA DRAAF is a two-year collaborative effort with University of Caen (F. Jurie) and the industrial partner EVITECH (P. Bernas) focused on modelling and recognition of violent behaviour in surveillance videos. The project aims to develop image recognition models and algorithms to automatically detect weapons, gestures and actions using recent advances in computer vision and deep learning to provide an affordable real-time solution reducing effects of threats in public places.

9.2. European Initiatives

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

Participant: Ivan Laptev.

WILLOW will be funded in part from 2013 to 2018 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.’

9.2.2. 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. IMPACT: Intelligent machine perception

Participants: Josef Sivic, Jean Ponce, Ivan Laptev.

IMPACT is a 5-year collaborative project with Czech Technical University, Center for Robotics, Informatics and Cybernetics (CIIRC) (2017-2022). The IMPACT project focuses on fundamental and applied research in computer vision, machine learning and robotics to develop machines that learn to perceive, reason, navigate and interact with complex dynamic environments. For example, people easily learn how to change a flat tire of a car or perform resuscitation by observing other people doing the same task. This involves advanced visual intelligence abilities such as interpreting sequences of human actions that manipulate objects to achieve a specific task. Currently, however, there is no artificial system with a similar level of cognitive visual competence. Breakthrough progress in intelligent machine perception will have profound implications on our everyday lives as well as science and commerce, with smart assistive robots that automatically learn new skills from the Internet, safer cars that autonomously navigate in difficult changing conditions, or intelligent glasses that help people navigate never seen before environments.

9.3.2. Associate team GAYA

Participants: Jean Ponce, Matthew Trager.

GAYA is a joint research team bringing together two Inria project-teams (Thoth, Grenoble and WILLOW, Paris) and Carnegie Mellon University, USA. It focuses on two research themes: (i) semantic structured interpretation of videos, and (ii) studying the geometric properties of object shapes to enhance state-of-the-art object recognition approaches.

Interpreting videos semantically in a general setting, involving various types of video content like home video clips, news broadcasts, feature films, which contain a lot of clutter, non-rigid motion, many “actors” performing actions, person-object and person-person interactions, varying viewpoints, is challenging. This task is being examined increasingly over the past decade, with the availability of large video resources, e.g., YouTube. Despite this progress, an effective video representation for recognizing actions is still missing. To address this critical challenge, we propose a joint optimization framework, wherein we learn the video representation and also develop models for action recognition. Specifically, we aim to exploit the spatio-temporal relations among pixels in a video through graphical models and novel deep learning feature representations.

The second research theme explores geometric aspects of computer vision, in particular how to model three-dimensional objects from their two-dimensional projections, and how the appearance of these objects evolves with changes in viewpoint. Beyond its theoretical interest, this work is critical for developing object recognition algorithms that take into account the three-dimensional nature of the visual world and go beyond the template-matching approaches dominant today. Duality is an important concept in this area, and we are investigating its application to the construction of visual hulls as well as the characterization of the topology of image contours using the Gauss map. Existing results are essentially limited to the Euclidean setting, and we are investigating their generalization to the general projective case.

Partners: CMU (Deva Ramanan, Martial Hebert, Abhinav Gupta, Gunnar Sigurdsson), Inria Thoth (Cordelia Schmid, Karteek Alahari, Pavel Tokmakov).

9.4. International Research Visitors

9.4.1. Visits of International Scientists

Alexei Efros (Professor, UC Berkeley, USA) visited Willow during May-June. Ramazan Cinbis (Middle East Technical University) and David Fouhey (University of Michigan) visited Willow in July-August and September-November, respectively. Akihiko Torii (Tokyo Institute of Technology) spent sabbatical at Willow from Apr to August 2018. Finally, Pierre-Yves Masse (post-doc, Czech Technical University) spent 50% of his time at Sierra (F. Bach) and Willow teams as a visiting post-doc within the framework of collaboration with the Intelligent Machine Perception project lead by J. Sivic at the 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 was program co-chair of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

10.1.1.2. Member of the Organizing Committees

- G. Varol and Y. Hasson are co-organizers of “Women in Computer Vision Workshop” at European Conference on Computer Vision (ECCV), 2018.
- I. Laptev and J. Sivic were co-organizers of “Fine-grained instructional video understanding workshop" at IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

10.1.2. Scientific Events Selection

10.1.2.1. Area chairs

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018 (J. Sivic).
- European Conference on Computer Vision (ECCV), 2018 (I. Laptev, J. Sivic).

10.1.2.2. Member of the Conference Program Committees / reviewer

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018 (J.-B. Alayrac, M. Oquab, R. Rezende, I. Rocco, G. Varol).
- European Conference on Computer Vision (ECCV), 2018 (A. Miech, G. Varol, I. Rocco, S. Zagaryuko).
- Neural Information Processing Systems (NIPS), 2018 (J. Sivic).
- Asian Conference on Computer Vision (ACCV), 2018 (G. Varol).

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- 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).

10.1.3.2. Reviewer - Reviewing Activities

- International Journal of Computer Vision (G. Cheron, M. Trager, G. Varol).
- IEEE Transactions on Pattern Analysis and Machine Intelligence (J.-B. Alayrac, G. Cheron, M. Trager, G. Varol).
- IEEE Transactions on Circuits and Systems for Video Technology (G. Varol).

10.1.4. Other

- J. Sivic is senior fellow of the Neural Computation and Adaptive Perception program of the Canadian Institute of Advanced Research.

10.1.5. Invited Talks

- I. Laptev, Keynote, ICCVG, Warsaw, September, 2018.
- I. Laptev, Invited talk, EPIC Workshop in conjunction with ECCV’18, Munich, September 2018.
- I. Laptev, Tutorial, BMVC’18, Newcastle, September 2018.

- I. Laptev, Invited talk, Workshop on Brave New Ideas for Video Understanding, in conjunction with CVPR'18, Salt Lake City, June 2018.
- I. Laptev, Invited talk, Journee AI, SAFRAN, Paris, June 2018.
- I. Laptev, Invited talk, Integrating Vision and Language, Tartu, March 2018.
- I. Laptev, Keynote, 36th Annual Swedish Symposium on Image Analysis, Stockholm, March 2018.
- A. Miech, Invited talk, LSCP-ENS seminar, Paris, March 2018.
- A. Miech, Invited talk, Google, Paris, July 2018.
- A. Miech, Invited talk, Google, Mountain View, July 2018.
- A. Miech, Invited talk, Paris ML Meetup, University of Bristol Computer Vision Seminar, Bristol, UK, November 2018.
- J. Ponce, Artificial Intelligence, French-American Joint Committee Meeting on Science and Technology, College de France, March 9, 2018.
- J. Ponce, Weakly supervised structure discovery in images and videos, Intelligent robots: autonomy and vision, NYU Abu Dhabi, March 13, 2018.
- J. Ponce, From vision and robotics to artificial intelligence, Robotics AI: Data science vs motion intelligence symposium co-organized by the French and German Academies of Sciences, Sep. 5, 2018.
- J. Ponce, Shape, contours, cameras and eyes, Workshop in honor of Jan Koenderink, UC Berkeley, UC berkeley, Oct. 24.
- J. Ponce, Weakly supervised structure discovery in images and videos, NYU Tandon School of Engineering, New York, Nov. 2, 2018.
- J. Ponce, Computer vision and visual recognition: historical perspective, new results and challenges, University of Zagreb, Zagreb, Croatia, Nov. 9, 2018.
- J. Sivic, Invited talk, Paris Sciences et Data, PSL, 02/2018.
- J. Sivic, Invited talk, AIME@CZ - Czech workshop on applied mathematics in engineering, Czech Technical University, 02/2018.
- J. Sivic, Invited talk, Deep Learning Workshop, CVPR 2018, Salt Lake City, June 2018.
- J. Sivic, Invited talk, Landmark Recognition Workshop, CVPR 2018, Salt Lake City, June 2018.
- J. Sivic, Seminar, UC Berkeley, June 2018.
- J. Sivic, Invited talk, Workshop on YouTube-8M Large-Scale Video Understanding, ECCV 2018, Munich, September 2018.
- J. Sivic, Invited talk, Prague Informatics Seminar, Charles University, Prague, April 2018.
- G. Varol, Invited talk, MPI for Informatics, Saarbrucken, Germany 11/2018.
- G. Varol, Invited talk, Istanbul Technical University, Istanbul, Turkey 10/2018.
- G. Varol, Invited talk, Deep Learning Meetup at Station F, Paris, France 09/2018.
- G. Varol, Invited talk, BNP Paribas - Prairie Summer School, Paris, France 08/2018.
- G. Varol, Invited talk, CTU Center for Machine Perception, Prague, Czech Republic 06/2018.
- G. Varol, Invited talk, MPI for Intelligent Systems, Tubingen, Germany 04/2018.

10.1.6. Leadership within the Scientific Community

- Member of the advisory board for the IBM Watson AI Xprize (J. Ponce).
- Member of the steering committee of France AI (J. Ponce).
- Member, advisory board, Computer Vision Foundation (J. Sivic).

10.1.7. Scientific Expertise

- J. Ponce, coordinator of the AI theme for the joint French-American Committee on Science and Technology, 2018-.

10.1.8. Research Administration

- Member, Bureau du comité des projets, Inria, Paris (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), 2012-2018 (J. Sivic).
- Member of the Hiring Committe for the tenure track position at CentraleSupelec (I. Laptev).
- Member of the Hiring Committee for Professor of Computer Vision at CentraleSupelec (I. Laptev).

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

- Master : M. Aubry, K. Alahari, I. Laptev and J. Sivic "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.
- J. Ponce co-organized the PRAIRIE AI Summer School, Grenoble, 2018, which brought together 200 participants representing 44 different nationalities, and selected from 700 applications, with 60% students, 15% academics , and 25% industrials. 25% of these participants were women.

10.2.2. Supervision

PhD : Jean-Baptiste Alayrac, “Structured Learning from Videos and Language”, graduated in September 2018, I. Laptev, J. Sivic and S. Lacoste-Julien (Inria SIERRA / U. Montreal).

PhD : Guilhem Cheron, “Structured modeling and recognition of human actions in video”, graduated in Dec 2018, I. Laptev and C. Schmid.

PhD : Maxime Oquab, “Convolutional neural networks: towards less supervision for visual recognition”, defended on 26 January 2018, L. Bottou (Facebook AI Research), I. Laptev and J. Sivic.

PhD : Matthew Trager, “Cameras, Shapes, and Contours: Geometric Models in Computer Vision”, graduated in 2018, J. Ponce and M. Hebert (CMU).

PhD : Tuang Hung VU, “Learning visual models for person detection and action prediction”, graduated in 2018 , I. Laptev.

PhD in progress : Vo Van Huy, started in Dec 2018, J. Ponce.

PhD in progress : Robin Strudel, “Learning and transferring complex robot skills from human demonstrations”, started in Oct 2018, I. Laptev, C. Schmid and J. Sivic.

PhD in progress : Yann Labbe, “Generalizing robotic sensorimotor skills to new tasks and environments", started in Oct 2018, J. Sivic and I. Laptev.

PhD in progress : Thomas Eboli, started in Oct 2017, J. Ponce.

PhD in progress : Zongmian Li, “Learning to manipulate objects from instructional videos”, started in Oct 2017, I. Laptev, J. Sivic and N. Mansard (LAAS/CNRS, Toulouse).

PhD in progress : Yana Hasson, started in Nov 2017, I. Laptev and C. Schmid.

PhD in progress : Dmitry Zhukov, “Learning from instruction videos for personal assistants”, started in Oct 2017, 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", 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 : Theophile Dalens, "Learning to analyze and reconstruct architectural scenes", starting in Jan 2015, M. Aubry and J. Sivic.

10.2.3. Juries

PhD thesis committee:

- Taylor MORDAN, Sorbonne Universite, France, 2018, (J. Sivic, rapporteur).
- Stephane LATHUILIERE, Universite Grenoble Alpes, France, 2018, (J. Sivic, rapporteur).
- Tuan-Hung VU, PSL University, France, 2018, (J. Sivic, examinateur).
- Siddhartha Chandra, CentraleSupelec, France, 2018, (I. Laptev examinateur).
- Sergey Zagoruyko, l?Universite Paris-Est, France, 2018, (I. Laptev examinateur).
- Joris Guerin, Arts et Metiers ParisTech, France, 2018, (I. Laptev examinateur).
- Guilhem Cheron, PSL, France, 2018, (J. Ponce examinateur).
- Pavel Tokmakov, Universite Grenoble Alpes, France, 2018, (J. Ponce examinateur).

10.3. Popularization

10.3.1. Articles and contents

- J. Ponce was the subject of an article in the photography magazine Polka. He was also interviewed by France Culture, Le Monde, Paris-Match, Québec Science, and Science et Vie.
- J. Ponce was interviewed by the Académie des Sciences working group on Artificial Intelligence on Oct. 2, 2018.

10.3.2. Interventions

- J. Ponce participated in round tables about AI at the France Culture Forum in Paris, March 1, 2018, at the BNP Parisbas summer school Aug. 23, 2018, at the French embassy in Berlin on Nov. 6, 2018, at the Erasme-Descartes conference in Paris on Nov. 15, 2018, and at the AI summit in New York City on Dec. 6, 2018.
- J. Ponce gave general audience lectures at the X-IA meeting in Palaiseau, Oct. 8, 2018 and at the Institut français de Croatie in Zagreb, Croatia, Nov. 9, 2018.

10.3.3. Creation of media or tools for science outreach

- I. Laptev, Y. Hasson, S. Allayen, T. Eboli, I. Kalyevath, Y. Labbe, R. Strudel, G. Varol, D. Zhukov, Presentation of computer vision for high-school students, Inria, December 2018.

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] J.-B. ALAYRAC. *Structured Learning from Videos and Language*, Ecole normale supérieure - ENS PARIS, September 2018, https://hal.inria.fr/tel-01885412

[2] G. Chéron. *Structured modeling and recognition of human actions in video*, Ecole normale supérieure - ENS PARIS, December 2018, https://hal.inria.fr/tel-01975247

[3] M. Oquab. *Convolutional neural networks: towards less supervision for visual recognition*, Ecole Normale Supérieure (ENS) ; ED 386 : École doctorale de sciences mathématiques de Paris centre, UPMC, January 2018, https://hal.inria.fr/tel-01803967

[4] M. Trager. *Cameras, Shapes, and Contours: Geometric Models in Computer Vision*, Ecole Normale Superieure de Paris - ENS Paris, July 2018, https://hal.inria.fr/tel-01856415

[5] T.-H. Vu. *Learning visual models for person detection and action prediction*, Ecole Normale Superieure de Paris - ENS Paris, September 2018, https://hal.inria.fr/tel-01861455

Articles in International Peer-Reviewed Journals

[6] B. Ham, M. Cho, J. Ponce. *Robust Guided Image Filtering Using Nonconvex Potentials*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", January 2018, vol. Vol. 40, n^{o} No. 1, pp. 291-207, Accepted pending minor revision [*DOI :* 10.1109/TPAMI.2017.2669034], https://hal.archives-ouvertes.fr/hal-01279857

[7] B. Ham, M. Cho, C. Schmid, J. Ponce. *Proposal Flow: Semantic Correspondences from Object Proposals*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", July 2018, vol. 40, n^{o} 7, pp. 1711-1725 [*DOI :* 10.1109/TPAMI.2017.2724510], https://hal.inria.fr/hal-01644132

[8] K. Kohn, B. Sturmfels, M. Trager. *Changing Views on Curves and Surfaces*, in "Acta Mathematica Vietnamica", 2018, https://arxiv.org/abs/1707.01877 - 31 pages [*DOI :* 10.1007/s40306-017-0240-1], https://hal.inria.fr/hal-01676208

[9] I. Rocco, R. Arandjelović, J. Sivic. *Convolutional neural network architecture for geometric matching*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2018, pp. 1-14 [*DOI :* 10.1109/TPAMI.2018.2865351], https://hal.archives-ouvertes.fr/hal-01859616

[10] G. Varol, I. Laptev, C. Schmid. *Long-term Temporal Convolutions for Action Recognition*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", June 2018, vol. 40, n^{o} 6, pp. 1510-1517, https://arxiv.org/abs/1604.04494 [*DOI :* 10.1109/TPAMI.2017.2712608], https://hal.inria.fr/hal-01241518

[11] Y. Zhang, Y. Su, J. Yang, J. Ponce, H. Kong. *When Dijkstra meets vanishing point: a stereo vision approach for road detection*, in "IEEE Transactions on Image Processing", 2018, pp. 1-12, https://hal.archives-ouvertes.fr/hal-01678548

International Conferences with Proceedings

[12] B. Bukh, X. Goaoc, A. Hubard, M. Trager. *Consistent Sets of Lines with no Colorful Incidence*, in "SoCG 2018 - 34thInternational Symposium on Computational Geometry", Budapest, Hungary, E. Speckmann, C. D. Tóth (editors), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, June 2018, pp. 1-20, https://arxiv.org/abs/1803.06267 , https://hal-upec-upem.archives-ouvertes.fr/hal-01744125

[13] J. Carpentier, G. Saurel, G. Buondonno, J. Mirabel, F. Lamiraux, O. Stasse, N. Mansard. *The Pinocchio C++ library – A fast and flexible implementation of rigid body dynamics algorithms and their*

analytical derivatives, in "SII 2019 - International Symposium on System Integrations", Paris, France, January 2019, https://hal.laas.fr/hal-01866228

[14] G. Chéron, J.-B. Alayrac, I. Laptev, C. Schmid. *A flexible model for training action localization with varying levels of supervision*, in "NIPS 2018 - 32nd Conference on Neural Information Processing Systems", Montréal, Canada, December 2018, pp. 1-17, https://arxiv.org/abs/1806.11328 , https://hal.inria.fr/hal-01937002

[15] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, B. Russell. *Localizing Moments in Video with Temporal Language*, in "Empirical Methods in Natural Language Processing (EMNLP)", Brussels, Belgium, October 2018, https://arxiv.org/abs/1809.01337 - EMNLP 2018, https://hal.archives-ouvertes.fr/hal-01976945

[16] E. Oyallon, E. Belilovsky, S. Zagoruyko, M. Valko. *Compressing the Input for CNNs with the First-Order Scattering Transform*, in "European Conference on Computer Vision", Munich, Germany, 2018, https://hal.inria.fr/hal-01850921

[17] I. Rocco, R. Arandjelović, J. Sivic. *End-to-end weakly-supervised semantic alignment*, in "CVPR 2018 - IEEE Conference on Computer Vision and Pattern Recognition", Salt Lake City, UT, United States, June 2018, pp. 1-9, https://hal.archives-ouvertes.fr/hal-01859628

[18] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, J. Sivic. *Neighbourhood Consensus Networks*, in "32nd Conference on Neural Information Processing Systems (NIPS 2018)", Montréal, Canada, December 2018, https://arxiv.org/abs/1810.10510 , https://hal.archives-ouvertes.fr/hal-01905474

[19] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, T. Pajdla. *Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions*, in "IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)", Salt Lake City, UT, United States, June 2018, https://arxiv.org/abs/1707.09092 , https://hal.archives-ouvertes.fr/hal-01859660

[20] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, A. Torii. *InLoc: Indoor Visual Localization with Dense Matching and View Synthesis*, in "CVPR 2018 - IEEE Conference on Computer Vision and Pattern Recognition", Salt Lake City, United States, June 2018, https://arxiv.org/abs/1803.10368 , https://hal.archives-ouvertes.fr/hal-01859637

[21] M. Trager, B. Osserman, J. Ponce. *On the Solvability of Viewing Graphs*, in "European Conference on Computer Vision 2018 (ECCV 2018)", Munich, Germany, September 2018, https://hal.inria.fr/hal-01856159

[22] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, C. Schmid. *BodyNet: Volumetric Inference of 3D Human Body Shapes*, in "ECCV 2018 - 15th European Conference on Computer Vision", Munich, Germany, September 2018, pp. 1-27, https://arxiv.org/abs/1804.04875 , https://hal.inria.fr/hal-01852169

Other Publications

[23] R. Budhiraja, J. Carpentier, N. Mansard. *Dynamics Consensus between Centroidal and Whole-Body Models for Locomotion of Legged Robots*, September 2018, Accepted for IEEE International Conference on Robotics and Automation 2019, https://hal.laas.fr/hal-01875031

[24] G. Chéron, A. Osokin, I. Laptev, C. Schmid. *Modeling Spatio-Temporal Human Track Structure for Action Localization*, January 2019, https://arxiv.org/abs/1806.11008 - working paper or preprint, https://hal.inria.fr/hal-01979583

[25] T. Eboli, J. Sun, J. Ponce. *Neural Embedding of an Iterative Deconvolution Algorithm for Motion Blur Estimation and Removal*, August 2018, working paper or preprint, https://hal.inria.fr/hal-01857177

[26] M. Hahn, N. Ruiz, J.-B. Alayrac, I. Laptev, J. M. Rehg. *Learning to Localize and Align Fine-Grained Actions to Sparse Instructions*, January 2019, https://arxiv.org/abs/1809.08381 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01979719

[27] X. S. Hu, S. Zagoruyko, N. Komodakis. *Exploring Weight Symmetry in Deep Neural Networks*, December 2018, https://arxiv.org/abs/1812.11027 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01978633

[28] B. Kim, J. Ponce, B. Ham. *Deformable Kernel Networks for Joint Image Filtering*, October 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01857016

[29] A. Miech, I. Laptev, J. Sivic. *Learning a Text-Video Embedding from Incomplete and Heterogeneous Data*, January 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01975102

[30] B. Osserman, M. Trager. *Multigraded Cayley-Chow Forms*, August 2018, working paper or preprint, https://hal.inria.fr/hal-01856190

[31] J. Peyre, I. Laptev, C. Schmid, J. Sivic. *Detecting rare visual relations using analogies*, January 2019, https://arxiv.org/abs/1812.05736 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01975760

[32] M. Trager, J. Ponce. *In Defense of Relative Multi-View Geometry*, August 2018, working paper or preprint, https://hal.inria.fr/hal-01676732

[33] T.-H. Vu, A. Osokin, I. Laptev. *Tube-CNN: Modeling temporal evolution of appearance for object detection in video*, January 2019, https://arxiv.org/abs/1812.02619 - 13 pages, 8 figures, technical report, https://hal.archives-ouvertes.fr/hal-01980339