Activity Report 2019

Team WILLOW

Models of visual object recognition and scene understanding

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions).

RESEARCH CENTER
Paris

THEME
Vision, perception and multimedia interpretation

Table of contents

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]
Justin Carpentier [Inria, Researcher, from Oct 2019]
Ivan Laptev [Inria, Senior Researcher, HDR]
Jean-Paul Laumond [CNRS, Senior Researcher, from Feb 2019, HDR]
Josef Sivic [Inria, Senior Researcher, HDR]

Post-Doctoral Fellows

Vijaya Kumar Reddy [Inria]
Makarand Tapaswi [Inria]
Sergey Zagoruyko [Inria, until Feb 2019]

PhD Students

Minttu Alakuijala [Inria, from Feb 2019]
Thomas Eboli [École Normale Supérieure de Paris]
Aamr El Kazdadi [Inria, from Oct 2019]
Pierre-Louis Guhur [Univ Paris-Saclay, from Sep 2019]
Yana Hasson [Inria]
Yann Labbe [École Normale Supérieure de Cachan]
Bruno Lecouat [Inria, from Sep 2019]
Zongmian Li [Inria]
Antoine Miech [Inria]
Julia Peyre [Inria, until Aug 2019]
Ronan Riochet [Inria]
Ignacio Rocco Spremolla [Inria]
Robin Strudel [École Normale Supérieure de Paris]
Gul Varol Simsekli [Inria, until Jun 2019]
Van Huy Vo [École Normale Supérieure de Paris]
Dimitri Zhukov [Inria]

Technical staff

Sofiane Allayen [Inria, until Apr 2019]
Yann Dubois de Mont Marin [Inria, from Dec 2019]
Igor Kalevatykh [Inria]
Oumayma Bounou [Inria, from Sep 2019]

Administrative Assistants
Helene Milome [Inria, until Jul 2019]
Mathieu Mourey [Inria, from Jul 2019]

Visiting Scientists
Pierre Yves Masse [Czech Technical University]
Vladimir Petrik [Czech Technical University]
Hazel Doughty [University of Bristol]

External Collaborator
Mathieu Aubry [École Nationale des Ponts et Chaussées]

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 Justin Carpentier as an Inria researcher. Jean-Paul Laumond has joined Willow as a senior CNRS researcher. Both Justin and Jean-Paul have robotics background and will strengthen the team in its new efforts towards learning embodied representations and robotics. Vladimir Petrik and Pierre-Yves Masse have been visiting post-docs within the framework of collaboration with the Intelligent Machine Perception project lead by J. Sivic at the Czech Technical University in Prague. Hazel Doughty has been a visiting PhD student from University of Bristol. Minttu Alakuijala, Aamr El Kazdadi, Pierre-Louis Guhur and Bruno Lecouat have joined Willow as new PhD students.

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.

3.5. Learning embodied representations

Computer vision has come a long way toward understanding images and videos in terms of scene geometry, object labels, locations and poses of people or classes of human actions. This “understanding”, however, remains largely disconnected from reasoning about the physical world. For example, what will happen if removing a tablecloth from a setted table? What actions will be needed to resume an interrupted meal? We believe that a true *embodied* understanding of dynamic scenes from visual observations is the next major research challenge. We plan to address this challenge by developing new models and algorithms with an emphasis on the synergy between vision, learning, robotics and natural language understanding. If successful, this research direction will bring significant advances in high-impact applications such as autonomous driving, home robotics and personal visual assistance.

Learning embodied representations is planned to be a major research axis for the successor of the Willow team. Meanwhile we have already started work in this direction and report our first results in Section 7.5.

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

- Four Prairie chairs awarded to I. Laptev, J.-P. Laumond, J. Ponce and J. Sivic by the international selection committee.
- Best Paper Award at FG (Automatic Face and Gesture Recognition - http://fg2019.org/awards/) 2019. (M. Tapaswi)
- Best paper finalist at CVPR 2019 for the work of Z. Li, J. Sedlar, J. Carpentier, I. Laptev, N. Mansard and J. Sivic Estimating 3D Motion and Forces of Person-Object Interactions From Monocular Video, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) [2]
- Best student paper awarded to H Cisneros, J Sivic, T Mikolov, Evolving Structures in Complex Systems at IEEE Symposium Series on Computational Intelligence (2019)

[2] More details at: https://www.ciirc.cvut.cz/vysledek-ciirc-cvut-se-dostal-do-uzsiho-vyberu-nejlepsich-clanku-prestizni-konference-cvpr-v-pocitacovem-videni/ and the list of all shortlisted papers is available at http://cvpr2019.thecvf.com/files/CVPR%202019%20-%20Welcome%20Slides%20Final.pdf

5.1.2. Visibility

- In 2019 we have recruited two excellent researchers with robotics background: Justin Carpentier and Jean-Paul Laumond, who will strengthen the team and will help developing a new research axis on learning embodied representations.
- J. Ponce co-organized the PRAIRIE AI Summer School, Paris, 2019.
- J. Ponce has been a key person in creating the PRAIRIE Institute for AI research in Paris, innagurated in October 2019.

BEST PAPER AWARD:

[11]
H. CISNEROS, J. SIVIC, T. MIKOLOV. *Evolving Structures in Complex Systems*, in "SSCI 2019 - IEEE Symposium Series on Computational Intelligence", Xiamen, China, December 2019, https://arxiv.org/abs/1911.01086 - IEEE Symposium Series on Computational Intelligence 2019 (IEEE SSCI 2019), https://hal.inria.fr/hal-02448134

6. New Software and Platforms

6.1. 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

6.2. VRAnalogy

Visual Relations detector using Analogy

KEYWORDS: Computer vision - Machine learning

FUNCTIONAL DESCRIPTION: Implementation of the paper "Detecting Unseen Visual Relations Using Analogies", Peyre et al', ICCV19

- Contact: Julia Peyre

6.3. d2-net

D2-Net: A Trainable CNN for Joint Description and Detection of Local Features

KEYWORD: Feature points

FUNCTIONAL DESCRIPTION: This repository contains the implementation of the following paper:

"D2-Net: A Trainable CNN for Joint Detection and Description of Local Features". M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler. CVPR 2019.

- Participants: Mihai Dusmanu, Ignacio Rocco Spremolla, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii and Torsten Sattler
- Contact: Ignacio Rocco Spremolla
- Publication: D2-Net: A Trainable CNN for Joint Detection and Description of Local Features
- URL: https://github.com/mihaidusmanu/d2-net

6.4. CrossTask

Cross-task weakly supervised learning from instructional videos

KEYWORDS: Videos - Machine learning

FUNCTIONAL DESCRIPTION: Open source release of the software package for the CVPR'19 paper "Cross-task weakly supervised learning from instructional videos" by D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev and J. Sivic

- Participants: Dimitri Zhukov, Jean-Baptiste Alayrac, Cinbis Gokberk, David Fouhey, Ivan Laptev and Josef Sivic
- Contact: Dimitri Zhukov
- URL: https://github.com/DmZhukov/CrossTask

6.5. MImE

Manipulation Imitation Environments

KEYWORDS: Robotics - Simulator - Computer vision

FUNCTIONAL DESCRIPTION: Simulation environment for learning robotics manipulation policies.

- Contact: Igor Kalevatykh
- URL: https://github.com/ikalevatykh/mime/

6.6. iReal

iReal: Implementing interactive scene understanding for a mixed reality device

KEYWORDS: Computer vision - Augmented reality - Demonstration

FUNCTIONAL DESCRIPTION: The goal of this project is to build a demonstration prototype of a smart personal assistant that sees and understands its surroundings to help the user navigate in unfamiliar environments, recognize new people and operate never seen before devices. The assistant will be implemented on the Microsoft HoloLens mixed reality device and will integrate the latest research software for automatic visual recognition and scene understanding developed in the Inria Willow team

- Contact: Mauricio Diaz

6.7. Sim2RealAugment

Learning to Augment Synthetic Images for Sim2Real Policy Transfer

KEYWORDS: Robotics - Computer vision

FUNCTIONAL DESCRIPTION: Implements the method described in Learning to Augment Synthetic Images for Sim2Real Policy Transfer (2019), A. Pashevich, R. Strudel, I. Kalevatykh, I. Laptev and C. Schmid, in Proc. IROS'19, Macau, China.

- Contact: Ivan Laptev
- URL: http://pascal.inrialpes.fr/data2/sim2real/

6.8. HowTo100M

HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips

KEYWORDS: Computer vision - Video analysis

FUNCTIONAL DESCRIPTION: Implements the method provides the dataset used in the paper HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips (2019), A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev and J. Sivic, in Proc. ICCV'19, Seoul, South Korea.

- Contact: Ivan Laptev
- URL: https://www.di.ens.fr/willow/research/howto100m/

6.9. ObMan

Learning joint reconstruction of hands and manipulated objects

KEYWORDS: Computer vision - 3D reconstruction

FUNCTIONAL DESCRIPTION: Implements the method and provides dataset used in the paper Learning joint reconstruction of hands and manipulated objects (2019), Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. Black, I. Laptev and C. Schmid, in Proc. CVPR'19, Long Beach, CA, USA.

- Contact: Yana Hasson
- URL: https://hassony2.github.io/obman.html

7. New Results

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

7.1.1. Learning joint reconstruction of hands and manipulated objects

Participants: Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael Black, Ivan Laptev, Cordelia Schmid.

Estimating hand-object manipulations is essential for interpreting and imitating human actions. Previous work has made significant progress towards reconstruction of hand poses and object shapes in isolation. Yet, reconstructing hands and objects during manipulation is a more challenging task due to significant occlusions of both the hand and object. While presenting challenges, manipulations may also simplify the problem since the physics of contact restricts the space of valid hand-object configurations. For example, during manipulation, the hand and object should be in contact but not interpenetrate. In [14] we regularize the joint reconstruction of hands and objects with manipulation constraints. We present an end-to-end learnable model that exploits a novel contact loss that favors physically plausible hand-object constellations. Our approach improves grasp quality metrics over baselines, using RGB images as input. To train and evaluate the model, we also propose a new large-scale synthetic dataset, ObMan, with hand-object manipulations. We demonstrate the transferability of ObMan-trained models to real data. Figure 1 presents some example results.

7.1.2. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features

Participants: Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, Torsten Sattler.

In [13], we address the problem of finding reliable pixel-level correspondences under difficult imaging conditions. We propose an approach where a single convolutional neural network plays a dual role: It is simultaneously a dense feature descriptor and a feature detector, as illustrated in Figure 2. By postponing the detection to a later stage, the obtained keypoints are more stable than their traditional counterparts based on early detection of low-level structures. We show that this model can be trained using pixel correspondences extracted from readily available large-scale SfM reconstructions, without any further annotations. The proposed method obtains state-of-the-art performance on both the difficult Aachen Day-Night localization dataset and the InLoc indoor localization benchmark, as well as competitive performance on other benchmarks for image matching and 3D reconstruction.

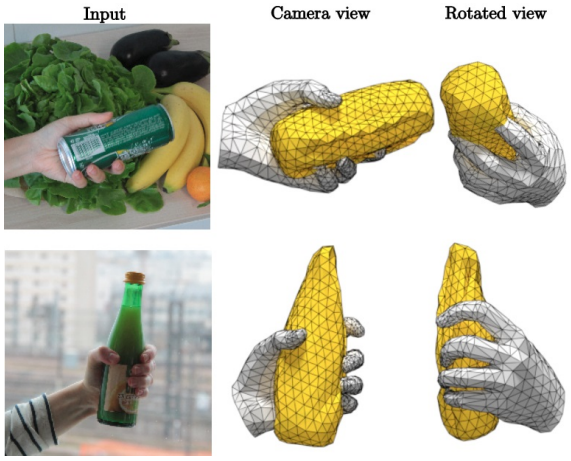

Figure 1. Our method jointly reconstructs hand and object meshes from a monocular RGB image. Note that the model generating the predictions for the above images, which we captured with an ordinary camera, was trained only on images from our synthetic dataset, ObMan.

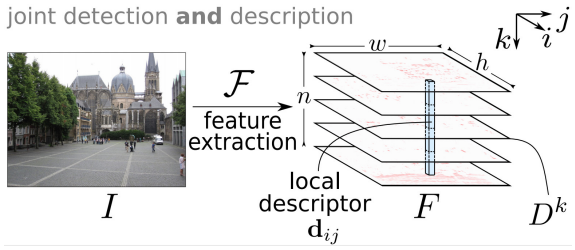

Figure 2. A feature extraction CNN $\mathcal{F}$ is used to extract feature maps that play a dual role: (i) local descriptors $\mathbf{d}_{ij}$ are simply obtained by traversing all the n feature maps D^k at a spatial position (i, j); (ii) detections are obtained by performing a non-local-maximum suppression on a feature map followed by a non-maximum suppression across each descriptor.

7.1.3. Is This The Right Place? Geometric-Semantic Pose Verification for Indoor Visual Localization

Participants: Hajime Taira, Ignacio Rocco, Jiri Sedlar, Masatoshi Okutomi, Josef Sivic, Tomas Pajdla, Torsten Sattler, Akihiko Torii.

Visual localization in large and complex indoor scenes, dominated by weakly textured rooms and repeating geometric patterns, is a challenging problem with high practical relevance for applications such as Augmented Reality and robotics. To handle the ambiguities arising in this scenario, a common strategy is, first, to generate multiple estimates for the camera pose from which a given query image was taken. The pose with the largest geometric consistency with the query image, e.g., in the form of an inlier count, is then selected in a second stage. While a significant amount of research has concentrated on the first stage, there is considerably less work on the second stage. In [21], we thus focus on pose verification. We show that combining different modalities, namely appearance, geometry, and semantics, considerably boosts pose verification and consequently pose accuracy, as illustrated in Figure 3. We develop multiple hand-crafted as well as a trainable approach to join into the geometric-semantic verification and show significant improvements over state-of-the-art on a very challenging indoor dataset.

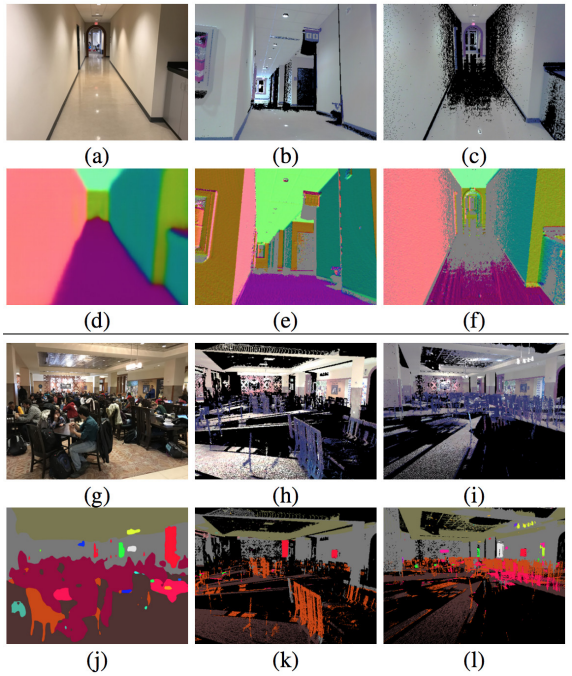

Figure 3. Given a set of camera pose estimates for a query image (a, g), we seek to identify the most accurate estimate. (b, h) Due to severe occlusion and weak textures, a state-of-the-art methodfails to identify the correct camera pose. To overcome those difficulties, we use several modalities along with visual appearance: (top) surface normals and (bottom) semantics. (c, i) Our approach verifies the estimated pose by comparing the semantics and surface normals extracted from the query (d, j) and database (f, l).

7.1.4. An Efficient Solution to the Homography-Based Relative Pose Problem With a Common Reference Direction

Participants: Yaqing Ding, Jian Yang, Jean Ponce, Hui Kong.

In [12], we propose a novel approach to two-view minimal-case relative pose problems based on homography with a common reference direction. We explore the rank-1 constraint on the difference between the Euclidean homography matrix and the corresponding rotation, and propose an efficient two-step solution for solving both the calibrated and partially calibrated (unknown focal length) problems. We derive new 3.5-point, 3.5-point, 4-point solvers for two cameras such that the two focal lengths are unknown but equal, one of them is unknown, and both are unknown and possibly different, respectively. We present detailed analyses and comparisons with existing 6-and 7-point solvers, including results with smart phone images.

7.1.5. Coordinate-Free Carlsson-Weinshall Duality and Relative Multi-ViewGeometry

Participants: Matthew Trager, Martial Hebert, Jean Ponce.

In [23], we present a coordinate-free description of Carlsson-Weinshall duality between scene points and camera pinholes and use it to derive a new characterization of primal/dual multi-view geometry. In the case of three views, a particular set of reduced trilinearities provide a novel parameterization of camera geometry that, unlike existing ones, is subject only to very simple internal constraints. These trilinearities lead to new "quasi-linear" algorithms for primal and dual structure from motion. We include some preliminary experiments with real and synthetic data.

7.1.6. Build your own hybrid thermal/EO camera for autonomous vehicle

Participants: Yigong Zhang, Yicheng Gao, Shuo Gu, Yubin Guo, Minghao Liu, Zezhou Sun, Zhixing Hou, Hang Yang, Ying Wang, Jian Yang, Jean Ponce, Hui Kong.

In [24], we propose a novel paradigm to design a hybrid thermal/EO (Electro-Optical or visible-light) camera, whose thermal and RGB frames are pixel-wisely aligned and temporally synchronized. Compared with the existing schemes, we innovate in three ways in order to make it more compact in dimension, and thus more practical and extendable for real-world applications. The first is a redesign of the structure layout of the thermal and EO cameras. The second is on obtaining a pixel-wise spatial registration of the thermal and RGB frames by a coarse mechanical adjustment and a fine alignment through a constant homography warping. The third innovation is on extending one single hybrid camera to a hybrid camera array, through which we can obtain wide-view spatially aligned thermal, RGB and disparity images simultaneously. The experimental results show that the average error of spatial-alignment of two image modalities can be less than one pixel. Some results of our method are illustrated in Figure 4.

7.2. Category-level object and scene recognition

7.2.1. Detecting unseen visual relations using analogies

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

In [19], 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 unseen at training. 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 em- beddings 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 three challenging datasets : on HICO-DET, our model achieves significant improvement over a strong baseline for both frequent and unseen triplets, and we observe similar improvement for the retrieval of unseen triplets with out-of- vocabulary predicates on the COCO-a dataset as well as the challenging unusual triplets in the UnRel dataset. Figure 5 presents an illustration of the approach.

7.2.2. SFNet: Learning Object-aware Semantic Correspondence

Participants: Junghyup Lee, Dohyung Kim, Jean Ponce, Bumsub Ham.

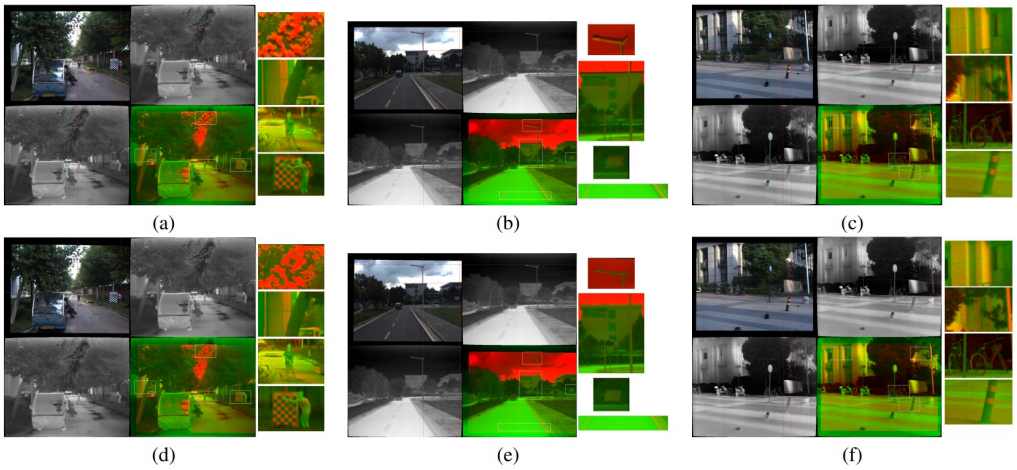

Figure 4. Results of alignment between the thermal and RGB frames of three sets of hybrid cameras before and after homography warping, respectively. (a), (b) and (c) are the alignment results before the homography warping, respectively. In each sub-figure, the layout of images is arranged as follows. Top-left: the aligned RGB image. Top-middle and bottom-left: the same aligned thermal image. Bottom-middle: the fusion image. (d), (e) and (f) are the alignment results after the homography warping, respectively. Likewise, the layout of images in each sub-figure is the same as those of (a), (b) and (c). To show the effect of homography rectification, we have overlaid red dotted lines horizontally and vertically onto the each sub-figure. In addition, the right column of each sub-figure zooms in four selected image regions to help us to view the warping result.

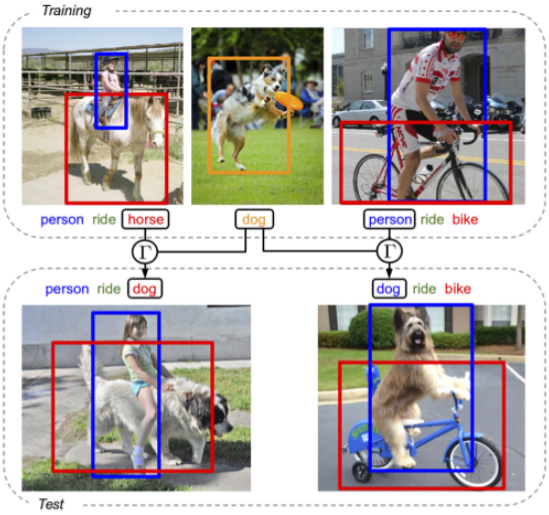

Figure 5. Illustration of transfer by analogy. We transfer visual representations of relations seen in the training set such as "person ride horse" to represent new unseen relations in the test set such as "person ride dog".

In [15], we address the problem of semantic correspondence, that is, establishing a dense flow field between images depicting different instances of the same object or scene category. We propose to use images annotated with binary foreground masks and subjected to synthetic geometric deformations to train a convolutional neural network (CNN) for this task. Using these masks as part of the supervisory signal offers a good compromise between semantic flow methods, where the amount of training data is limited by the cost of manually selecting point correspondences, and semantic alignment ones, where the regression of a single global geometric transformation between images may be sensitive to image-specific details such as background clutter. We propose a new CNN architecture, dubbed SFNet, which implements this idea. It leverages a new and differentiable version of the argmax function for end-to-end training, with a loss that combines mask and flow consistency with smoothness terms. Experimental results demonstrate the effectiveness of our approach, which significantly outperforms the state of the art on standard benchmarks. Figure 6 presents an illustration of the approach.

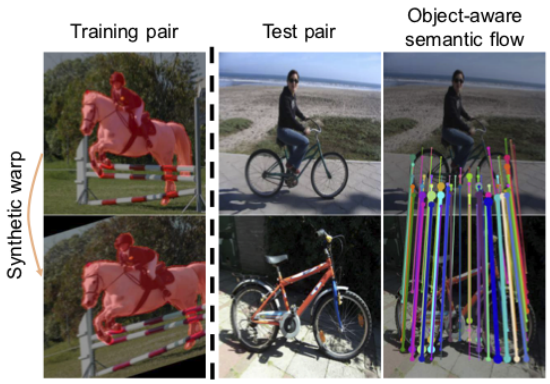

Figure 6. We use pairs of warped foreground masks obtained from a single image (left) as a supervisory signal to train our model.This allows us to establish object-aware semantic correspondences across images depicting different instances of the same object or scene category (right). No masks are required at test time.

7.2.3. Hyperpixel Flow: Semantic Correspondence with Multi-layer Neural Features

Participants: Juhong Min, Jongmin Kim, Jean Ponce, Minsu Cho.

In [17], we establish visual correspondences under large intra-class variations requires analyzing images at different levels , from features linked to semantics and context to local patterns, while being invariant to instance-specific details. To tackle these challenges, we represent images by "hyper-pixels" that leverage a small number of relevant features selected among early to late layers of a convolutional neural network. Taking advantage of the condensed features of hyperpixels, we develop an effective real-time matching algorithm based on Hough geometric voting. The proposed method, hyperpixel flow, sets a new state of the art on three standard benchmarks as well as a new dataset, SPair-71k, which contains a significantly larger number of image pairs than existing datasets, with more accurate and richer annotations for in-depth analysis. Figure 7 presents an illustration of the approach.

7.2.4. Exploring Weight Symmetry in Deep Neural Networks

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

In [7], 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,

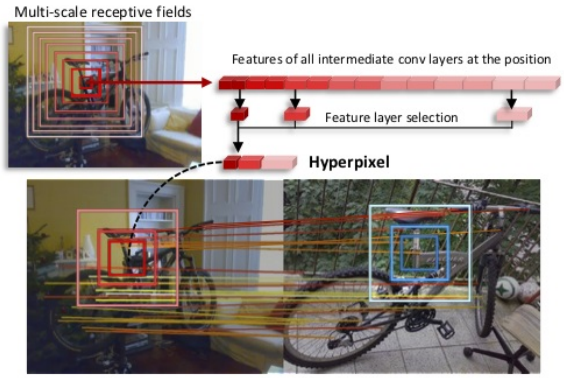

Figure 7. Hyperpixel flow. Top: The hyperpixel is a multi-layer pixel representation created with selected levels of features optimized for semantic correspondence. It provides multi-scale features, resolving local ambiguities. Bottom: The proposed method, hyperpixel flow, establishes dense correspondences in real time using hyperpixels.

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.

7.2.5. Bilinear image translation for temporal analysis of photo collections

Participants: Théophile Dalens, Mathieu Aubry, Josef Sivic.

In [5], we propose an approach for analyzing unpaired visual data annotated with time stamps by generating how images would have looked like if they were from different times. To isolate and transfer time dependent appearance variations, we introduce a new trainable bilinear factor separation module. We analyze its relation to classical factored representations and concatenation-based auto-encoders. We demonstrate this new module has clear advantages compared to standard concatenation when used in a bottleneck encoder-decoder convolutional neural network architecture. We also show that it can be inserted in a recent adversarial image translation architecture, enabling the image transformation to multiple different target time periods using a single network. We apply our model to a challenging collection of more than 13,000 cars manufactured between 1920 and 2000 and a dataset of high school yearbook portraits from 1930 to 2009, as illustrated in Figure 8. This allows us, for a given new input image, to generate a "history-lapse video" revealing changes over time by simply varying the target year. We show that by analyzing the generated history-lapse videos we can identify object deformations across time, extracting interesting changes in visual style over decades.

7.3. Image restoration, manipulation and enhancement

7.3.1. Deformable Kernel Networks for Joint Image Filtering

Participants: Beomjun Kim, Jean Ponce, Bumsub Ham.

Figure 8. Our method takes as input an image of an object (in green), such as a car, and generates what it would have looked like in another time-period (in blue). Each row shows temporal translation for a different input car image (in green). The translation model is trained on an unpaired dataset of cars with time stamps. We show that analyzing changes between the generated images reveal structural deformations in car shape and appearance over time.

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. In [28], 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 640×480. 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 3×3 kernels outperforms the state of the art by a significant margin.

7.3.2. Revisiting Non Local Sparse Models for Image Restoration

Participants: Bruno Lecouat, Jean Ponce, Julien Mairal.

In [29], we propose a differentiable algorithm for image restoration inspired by the success of sparse models and self-similarity priors for natural images. Our approach builds upon the concept of joint sparsity between groups of similar image patches, and we show how this simple idea can be implemented in a differentiable architecture, allowing end-to-end training. The algorithm has the advantage of being interpretable, performing sparse decompositions of image patches, while being more parameter efficient than recent deep learning methods. We evaluate our algorithm on grayscale and color denoising, where we achieve competitive results, and on demoisaicking, where we outperform the most recent state-of-the-art deep learning model with 47 times less parameters and a much shallower architecture. Figure 9 shows results of the proposed approach.

7.4. Human activity capture and classification

7.4.1. Video Face Clustering with Unknown Number of Clusters

Participants: Makarand Tapaswi, Marc T. Law, Sanja Fidler.

Figure 9. Demosaicking result obtained by our method. Top right: Ground truth. Middle: Image demosaicked with our sparse coding baseline without non-local prior. Bottom: demosaicking with sparse coding and non-local prior. The reconstruction does not exhibit any artefact on this image which is notoriously difficult for demosaicking.

Understanding videos such as TV series and movies requires analyzing who the characters are and what they are doing. We address the challenging problem of clustering face tracks based on their identity. Different from previous work in this area, we choose to operate in a realistic and difficult setting where: (i) the number of characters is not known a priori; and (ii) face tracks belonging to minor or background characters are not discarded.

To this end, we propose Ball Cluster Learning (BCL), a supervised approach to carve the embedding space into balls of equal size, one for each cluster (see Figure 10). The learned ball radius is easily translated to a stopping criterion for iterative merging algorithms. This gives BCL the ability to estimate the number of clusters as well as their assignment, achieving promising results on commonly used datasets. We also present a thorough discussion of how existing metric learning literature can be adapted for this task. This work has been published in [22].

7.4.2. Cross-task weakly supervised learning from instructional videos

Participants: Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, Josef Sivic.

In [25], we investigate learning visual models for the steps of ordinary tasks using weak supervision via instructional narrations and an ordered list of steps instead of strong supervision via temporal annotations. At the heart of our approach is the observation that weakly supervised learning may be easier if a model shares components while learning different steps: “pour egg” should be trained jointly with other tasks involving “pour” and “egg”. We formalize this in a component model for recognizing steps and a weakly supervised learning framework that can learn this model under temporal constraints from narration and the list of steps. Past data does not permit systematic studying of sharing and so we also gather a new dataset, CrossTask, aimed at assessing cross-task sharing. Our experiments demonstrate that sharing across tasks im- proves performance, especially when done at the component level and that our component model can parse previously unseen tasks by virtue of its compositionality. Figure 11 illustrates the idea of sharing step components between different tasks.

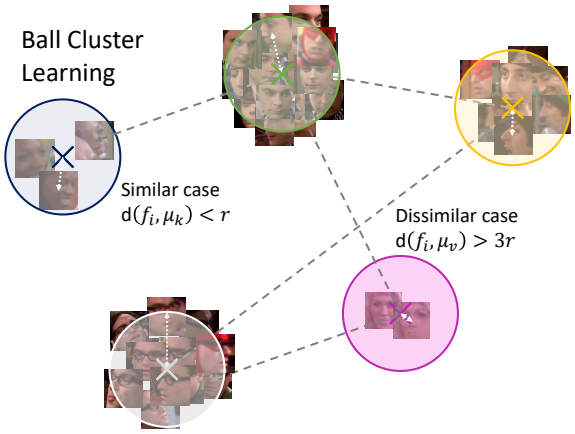

Figure 10. Ball Cluster Learning carves the feature space into balls of equal radius. The number of samples in the cluster does not affect the ball radius or minimum separation to other balls.

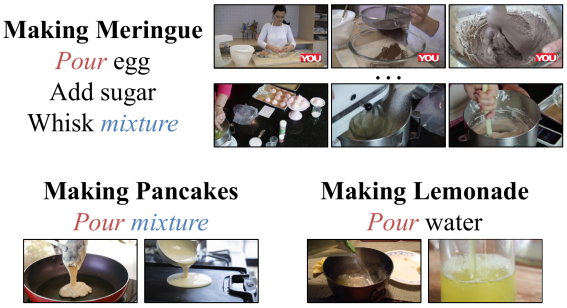

Figure 11. Our method begins with a collection of tasks, each con- sisting of an ordered list of steps and a set of instructional videos from YouTube. It automatically discovers both where the steps oc- cur and what they look like. To do this, it uses the order, narration and commonalities in appearance across tasks (e.g., the appear- ance of pour in both making pancakes and making meringue).

7.4.3. Leveraging the Present to Anticipate the Future in Videos

Participants: Antoine Miech, Ivan Laptev, Josef Sivic, Heng Wang, Lorenzo Torresani, Du Tran.

Anticipating actions before they are executed is crucial for a wide range of practical applications including autonomous driving and the moderation of live video streaming. While most prior work in this area requires partial observation of executed actions, in the paper we focus on anticipating actions seconds before they start (see Figure 12) Our proposed approach is the fusion of a purely anticipatory model with a complementary model constrained to reason about the present. In particular, the latter predicts present action and scene attributes, and reasons about how they evolve over time. By doing so, we aim at modeling action anticipation at a more conceptual level than directly predicting future actions. Our model outperforms previously reported methods on the EPIC-KITCHENS and Breakfast datasets. This paper was presented at the CVPR 2019 precognition workshop [34] and ranked second at the EPIC-KITCHENS action anticipation challenge.

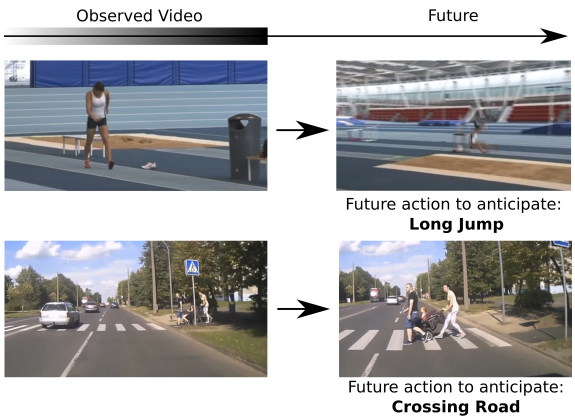

Figure 12. Examples of action anticipation in which the goal is to anticipate future actions in videos seconds before they are performed.

7.4.4. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips

Participants: Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, Josef Sivic.

Learning text-video embeddings usually requires a dataset of video clips with manually provided captions. However, such datasets are expensive and time consuming to create and therefore difficult to obtain on a large scale. In this work, we propose instead to learn such embeddings from video data with readily available natural language annotations in the form of automatically transcribed narrations (see Figure 13). The contributions of this work are three-fold. First, we introduce HowTo100M: a large-scale dataset of 136 million video clips sourced from 1.22M narrated instructional web videos depicting humans performing and describing over 23k different visual tasks. Our data collection procedure is fast, scalable and does not require any additional manual annotation. Second, we demonstrate that a text-video embedding trained on this data leads to state-of-the-art results for text-to-video retrieval and action localization on instructional video datasets such as YouCook2 or CrossTask. Finally, we show that this embedding transfers well to other domains: fine-tuning on generic Youtube videos (MSR-VTT dataset) and movies (LSMDC dataset) outperforms models trained on these datasets alone. This work was presented at ICCV 2019 [16].

Figure 13. We learn a joint text-video embedding by watching millions of narrated video clips of people performing diverse visual tasks. The learned embedding transfers well to other instructional and non-instructional text-video datasets.

7.4.5. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization?

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

Accurate visual localization is a key technology for autonomous navigation. 3D structure-based methods, as illustrated in Figure 14, employ 3D models of the scene to estimate the full 6 degree-of-freedom (DOF) pose of a camera very accurately. However, constructing (and extending) large-scale 3D models is still a significant challenge. In contrast, 2D image retrieval-based methods only require a database of geo-tagged images, which is trivial to construct and to maintain. They are often considered inaccurate since they only approximate the positions of the cameras. Yet, the exact camera pose can theoretically be recovered when enough relevant database images are retrieved. In [8], we demonstrate experimentally that large-scale 3D models are not strictly necessary for accurate visual localization. We create reference poses for a large and challenging urban dataset. Using these poses, we show that combining image-based methods with local reconstructions results in a higher pose accuracy compared to state-of-the-art structure-based methods, albeight at higher run-time costs. We show that some of these run-time costs can be alleviated by exploiting known database image poses. Our results suggest that we might want to reconsider the need for large-scale 3D models in favor of more local models, but also that further research is necessary to accelerate the local reconstruction process.

7.4.6. End-to-End Learning of Visual Representations from Uncurated Instructional Videos

Participants: Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, Andrew Zisserman.

Annotating videos is cumbersome, expensive and not scalable. Yet, many strong video models still rely on manually annotated data. With the recent introduction of the HowTo100M dataset, narrated videos now offer the possibility of learning video representations without manual supervision. In this work we propose a new learning approach, MIL-NCE, capable of addressing misalignments inherent to narrated videos (see Figure 15). With this approach we are able to learn strong video representations from scratch, without the need for any manual annotation. We evaluate our representations on a wide range of four downstream tasks over eight datasets: action recognition (HMDB-51, UCF-101, Kinetics-700), text-to-video retrieval (YouCook2, MSR-VTT), action localization (YouTube-8M Segments, CrossTask) and action segmentation (COIN). Our method outperforms all published self-supervised approaches for these tasks as well as several fully supervised baselines. This preprint [32] is currently under review.

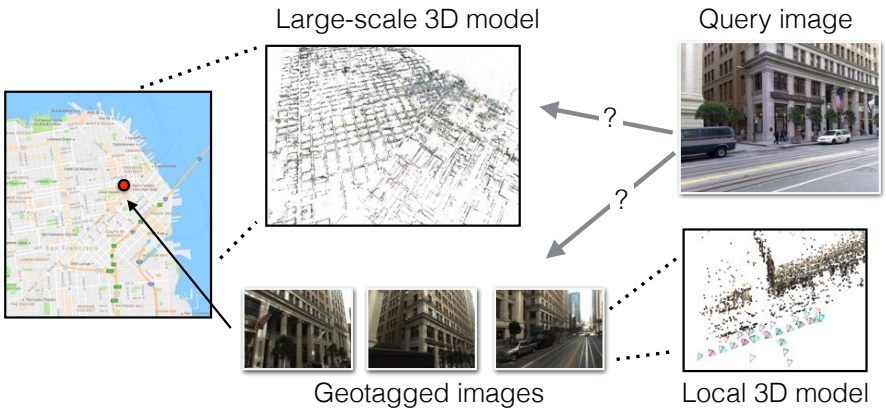

Figure 14. The state-of-the-art for large-scale visual localization. 2D image-based methods (bottom) use image retrieval and return the pose of the most relevant database image. 3D structure-based methods (top) use 2D-3D matches against a 3D model for camera pose estimation. Both approaches have been developed largely independently of each other and never compared properly before.

Figure 15. We describe an efficient approach to learn visual representations from highly misaligned and noisy narrations automatically extracted from instructional videos. Our video representations are learnt from scratch without relying on any manually annotated visual dataset yet outperform all self-supervised and many fully-supervised methods on several video recognition benchmarks.

7.4.7. Synthetic Humans for Action Recognition from Unseen Viewpoints

Participants: Gul Varol, Ivan Laptev, Cordelia Schmid, Andrew Zisserman.

In [35], the goal is to improve the performance of human action recognition for viewpoints unseen during training by using synthetic training data. Although synthetic data has been shown to be beneficial for tasks such as human pose estimation, its use for RGB human action recognition is relatively unexplored. We make use of the recent advances in monocular 3D human body reconstruction from real action sequences to automatically render synthetic training videos for the action labels. We make the following contributions: (i) we investigate the extent of variations and augmentations that are beneficial to improving performance at new viewpoints. We consider changes in body shape and clothing for individuals, as well as more action relevant augmentations such as non-uniform frame sampling, and interpolating between the motion of individuals performing the same action; (ii) We introduce a new dataset, SURREACT, that allows supervised training of spatio-temporal CNNs for action classification; (iii) We substantially improve the state-of-the-art action recognition performance on the NTU RGB+D and UESTC standard human action multi-view benchmarks; Finally, (iv) we extend the augmentation approach to in-the-wild videos from a subset of the Kinetics dataset to investigate the case when only one-shot training data is available, and demonstrate improvements in this case as well. Figure 16 presents an illustration of the approach.

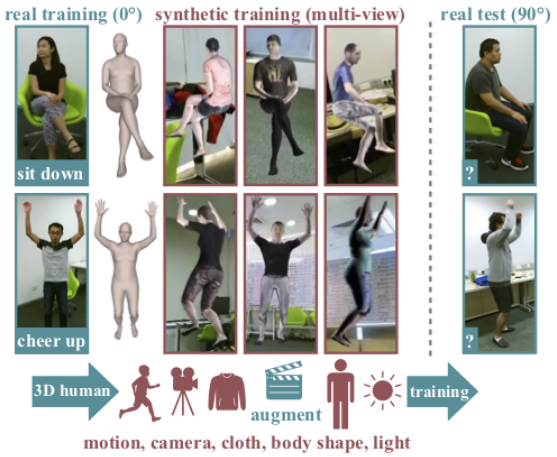

Figure 16. We estimate 3D shape from real videos and automatically render synthetic videos with action labels. We explore various augmentations for motions, viewpoints, and appearance. Training temporal CNNs with this data significantly improves the action recognition from unseen viewpoints.

7.5. Learning embodied representations and robotics

7.5.1. Roboticists and Reporters

Participants: Celine Pieters, Emmanuelle Danblon, Jean-Paul Laumond.

This paper reports on an experiment organized at the Cité des Sciences et de l'Industrie (CSI) of Paris in order to assess the importance of language in the representation and the integration of robots into the human culture. The experiment gathered specialized reporters and experts in robotics around a practical exercise of rhetoric. The objective of this work is to show that rhetoric is not a matter of communication, but a technique that allows to better understand the way roboticists understand their own discipline.

7.5.2. Robots

Participants: Jean-Paul Laumond, Denis Vidal.

What is a robot? How does it work? How is research progressing, what are the challenges and the economic and social questions posed by robotics in the twenty-first century? Today, as robots are becoming increasingly present in our professional, public and private lives, it is vital to understand their technological capabilities. We must more fully comprehend how they can help us and master their uses. Robots continue to fascinate us but our idea of them, stemming from literature and cinema, is often a purely imaginary one. This illustrated book accompanies the Robots exhibition at the Cité des sciences et de l'industrie. Figure 17 presents the front page of the exhibition.

Figure 17. Front page of the permanent exhibition at Cité des Sciences et de l'Industrie about Robotics.

7.5.3. Learning to Augment Synthetic Images for Sim2Real Policy Transfer

Participants: Alexander Pashevich, Robin Strudel, Igor Kalevatykh, Ivan Laptev, Cordelia Schmid.

Vision and learning have made significant progress that could improve robotics policies for complex tasks and environments. Learning deep neural networks for image understanding, however, requires large amounts of domain-specific visual data. While collecting such data from real robots is possible, such an approach limits the scalability as learning policies typically requires thousands of trials. In this paper [18], we attempt to learn manipulation policies in simulated environments. Simulators enable scalability and provide access to the underlying world state during training. Policies learned in simulators, however, do not transfer well to real scenes given the domain gap between real and synthetic data. We follow recent work on domain randomization and augment synthetic images with sequences of random transformations. Our main contribution is to optimize the augmentation strategy for sim2real transfer and to enable domain-independent policy learning. We design an efficient search for depth image augmentations using object localization as a proxy task. Given the resulting sequence of random transformations, we use it to augment synthetic depth images during policy learning. Our augmentation strategy is policy-independent and enables policy learning with no real images. We demonstrate our approach to significantly improve accuracy on three manipulation tasks evaluated on a real robot. Figure 18 presents an illustration of the approach.

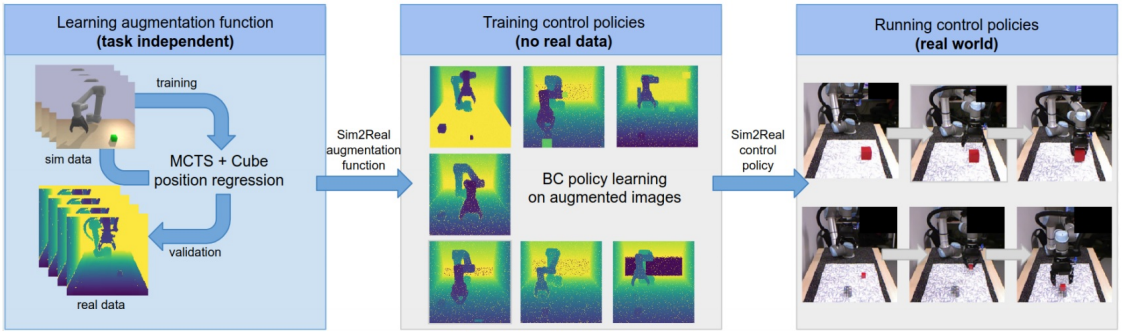

Figure 18. Overview of the method. Our contribution is the policy-independent learning of depth image augmentations (left). The resulting sequence of augmentations is applied to synthetic depth images while learning manipulation policies in a simulator (middle). The learned policies are directly applied to real robot scenes without finetuning on real images (right).

7.5.4. Learning to combine primitive skills: A step towards versatile robotic manipulation

Participants: Robin Strudel, Alexander Pashevich, Igor Kalevatykh, Ivan Laptev, Josef Sivic, Cordelia Schmid.

Manipulation tasks such as preparing a meal or assembling furniture remain highly challenging for robotics and vision. Traditional task and motion planning (TAMP) methods can solve complex tasks but require full state observability and are not adapted to dynamic scene changes. Recent learning methods can operate directly on visual inputs but typically require many demonstrations and/or task-specific reward engineering. In this paper [20], we aim to overcome previous limitations and propose a reinforcement learning (RL) approach to task planning that learns to combine primitive skills. First, compared to previous learning methods, our approach requires neither intermediate rewards nor complete task demonstrations during training. Second, we demonstrate the versatility of our vision-based task planning in challenging settings with temporary occlusions and dynamic scene changes. Third, we propose an efficient training of basic skills from few synthetic demonstrations by exploring recent CNN architectures and data augmentation. Notably, while all of our policies are learned on visual inputs in simulated environments, we demonstrate the successful transfer and high success rates when applying such policies to manipulation tasks on a real UR5 robotic arm. Figure 19 presents an illustration of the approach.

7.5.5. Monte-Carlo Tree Search for Efficient Visually Guided Rearrangement Planning

Participants: Sergey Zagoruyko, Yann Labbé, Igor Kalevatykh, Ivan Laptev, Justin Carpentier, Mathieu Aubry, Josef Sivic.

We address the problem of visually guided rearrangement planning with many movable objects, i.e., finding a sequence of actions to move a set of objects from an initial arrangement to a desired one, while relying on visual inputs coming from RGB camera. To do so, we introduce a complete pipeline relying on two key contributions. First, we introduce an efficient and scalable rearrangement planning method, based on a Monte-Carlo Tree Search exploration strategy. We demonstrate that because of its good trade-off between exploration and exploitation our method (i) scales well with the number of objects while (ii) finding solutions which require a smaller number of moves compared to the other state-of-the-art approaches. Note that on the contrary to many approaches, we do not require any buffer space to be available. Second, to precisely localize movable objects in the scene, we develop an integrated approach for robust multi-object workspace state estimation from a single uncalibrated RGB camera using a deep neural network trained only with synthetic data. We validate our multi-object visually guided manipulation pipeline with several experiments on a real

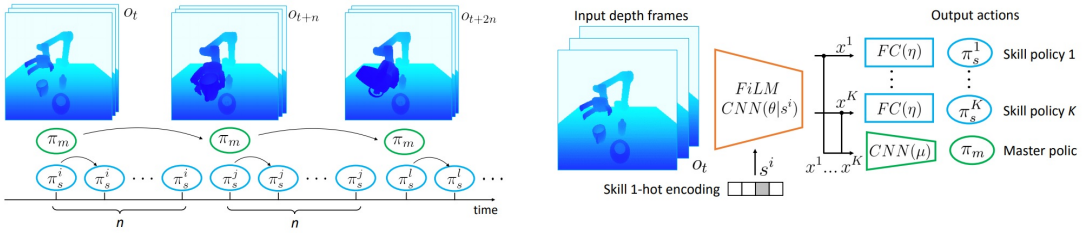

Figure 19. Illustration of our approach. (Left): Temporal hierarchy of master and skill policies. The master policy π_m is executed at a coarse interval of n time-steps to select among K skill policies $\pi_s^1 ... \pi_s^K$. Each skill policy generates control for a primitive action such as grasping or pouring. (Right): CNN architecture used for the skill and master policies.

UR-5 robotic arm by solving various rearrangement planning instances, requiring only 60 ms to compute the complete plan to rearrange 25 objects. In addition, we show that our system is insensitive to camera movements and can successfully recover from external perturbation. Figure 20 shows an example of the problems we consider. This work is under-review and an early pre-print is available [37].

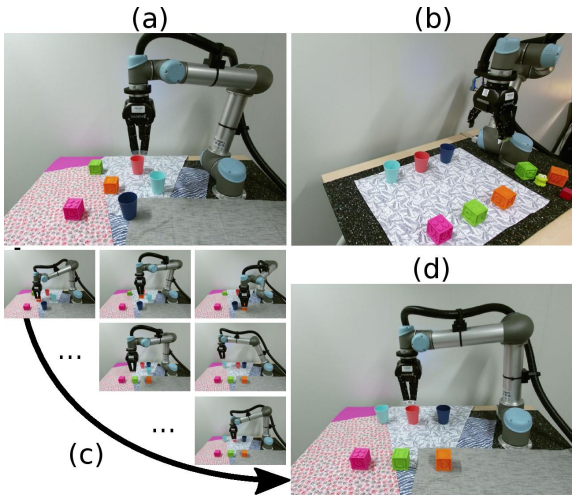

*Figure 20. **Visually guided rearrangement planning.** Given a source (a) and target (b) RGB images depicting a robot and multiple movable objects, our approach estimates the positions of objects in the scene without the need for explicit camera calibration and efficiently finds a sequence of robot actions (c) to re-arrange the scene into the target scene. Final object configuration after re-arrangement by the robot is shown in (d).*

7.5.6. Estimating the Center of Mass and the Angular Momentum Derivative for Legged Locomotion — A recursive approach

Participants: François Bailly, Justin Carpentier, Mehdi Benallegue, Bruno Watier, Philippe Soueres.

Estimating the center of mass position and the angular momentum derivative of legged systems is essential for both controlling legged robots and analyzing human motion. In this paper[4], a novel recursive approach to concurrently and accurately estimate these two quantities together is introduced. The proposed method employs kinetic and kinematic measurements from classic sensors available in robotics and biomechanics, to effectively exploits the accuracy of each measurement in the spectral domain. The soundness of the proposed approach is first validated on a simulated humanoid robot, where ground truth data is available, against an Extend Kalman Filter. The results demonstrate that the proposed method reduces the estimation error on the center of mass position with regard to kinematic estimation alone, whereas at the same time, it provides an accurate estimation of the derivative of angular momentum. Finally, the effectiveness of the proposed method is illustrated on real measurements, obtained from walking experiments with the HRP-2 humanoid robot. Figure 21 presents an illustration of the approach.

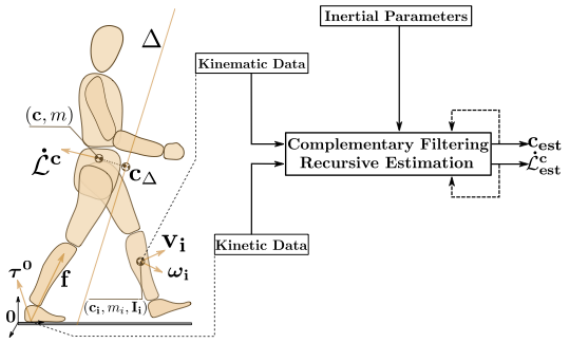

Figure 21. Illustration of the measurement apparatus. The several physical quantities involved in the estimation framework are displayed, as well as a simplified sketch of the estimation algorithm.

7.5.7. Dynamics Consensus between Centroidal and Whole-Body Models for Locomotion of Legged Robots

Participants: Rohan Budhiraja, Justin Carpentier, Nicolas Mansard.

It is nowadays well-established that locomotion can be written as a large and complex optimal control problem. Yet, current knowledge in numerical solver fails to directly solve it. A common approach is to cut the dimensionality by relying on reduced models (inverted pendulum, capture points, centroidal). However it is difficult both to account for whole-body constraints at the reduced level and also to define what is an acceptable trade-off at the whole-body level between tracking the reduced solution or searching for a new one. The main contribution of this paper [9] is to introduce a rigorous mathematical framework based on the Alternating Direction Method of Multipliers, to enforce the consensus between the centroidal state dynamics at reduced and whole-body level. We propose an exact splitting of the whole-body optimal control problem between the centroidal dynamics (under-actuation) and the manipulator dynamics (full actuation), corresponding to a rearrangement of the equations already stated in previous works. We then describe with details how alternating descent is a good solution to implement an effective locomotion solver. We validate this approach in simulation with walking experiments on the HRP-2 robot. Figure 22 presents a resulting motion of the proposed approach.

7.5.8. 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.

Figure 22. Walking sequence generated for HRP-2 robot using the proposed ADMM solver.

In this paper [10], 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. Figure 23 presents the logo of Pinocchio.

Figure 23. Logo of Pinocchio.

7.5.9. Crocoddyl: An Efficient and Versatile Framework for Multi-Contact Optimal Control

Participants: Carlos Mastalli, Rohan Budhiraja, Wolfgang Merkt, Guilhem Saurel, Bilal Hammoud, Maximilien Naveau, Justin Carpentier, Sethu Vijayakumar, Nicolas Mansard.

In this paper [31], we introduce Crocoddyl (Contact RObot COntrol by Differential DYnamic Library), an open-source framework tailored for efficient multi-contact optimal control. Crocoddyl efficiently computes the state trajectory and the control policy for a given predefined sequence of contacts. Its efficiency is due to the use of sparse analytical derivatives, exploitation of the problem structure, and data sharing. It employs differential geometry to properly describe the state of any geometrical system, e.g. floating-base systems. We have unified dynamics, costs, and constraints into a single concept-action-for greater efficiency and easy prototyping. Additionally, we propose a novel multiple-shooting method called Feasibility-prone Differential Dynamic Programming (FDDP). Our novel method shows a greater globalization strategy compared to classical Differential Dynamic Programming (DDP) algorithms, and it has similar numerical behavior to state-of-the-art multiple-shooting methods. However, our method does not increase the computational complexity typically encountered by adding extra variables to describe the gaps in the dynamics. Concretely, we propose two modifications to the classical DDP algorithm. First, the backward pass accepts infeasible state-control trajectories. Second, the rollout keeps the gaps open during the early "exploratory" iterations (as expected in multiple-shooting methods). We showcase the performance of our framework using different tasks. With our

method, we can compute highly-dynamic maneuvers for legged robots (e.g. jumping, front-flip) in the order of milliseconds. Figure 24 presents a resulting motion of the proposed approach.

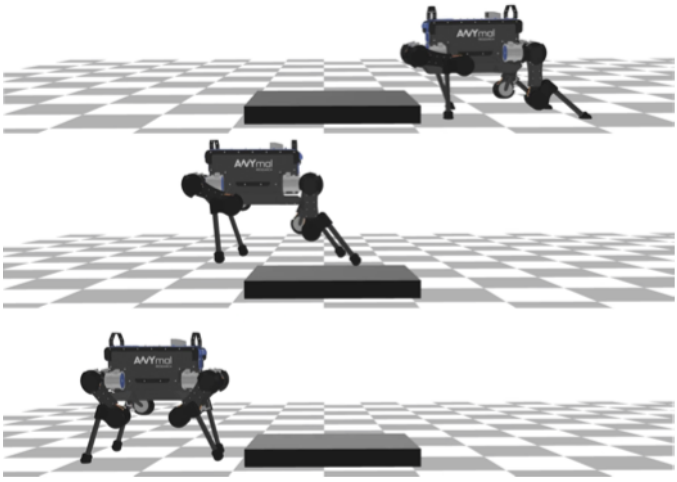

Figure 24. Crocoddyl: an efficient and versatile framework for multi-contact optimal control. Highly-dynamic maneuvers needed to traverse an obstacle with the ANYmal robot.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

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

Participants: Yana Hasson, Ivan Laptev, Jean Ponce, Josef Sivic, Dimitri Zhukov, Cordelia Schmid [Inria Thoth].

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: 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. PRAIRIE

Participants: Ivan Laptev, Jean-Paul Laumond, Jean Ponce, Josef Sivic.

The Prairie Institute (PaRis AI Research InstitutE) is one of the four French Institutes for Interdisciplinary Artificial Intelligence Research (3IA), which were created as part of the national French initiative on AI announced by President Emmanuel Macron on May 29, 2018. It brings together five academic partners (CNRS, Inria, Institut Pasteur, PSL University, and University of Paris) as well as 17 industrial partners, large corporations which are major players in AI at the French, European and international levels, as well as 45 Chair holders, including four of the members of WILLOW (Laumond, Laptev, Ponce, Sivic). Ponce is the scientific director of PRAIRIE.

9.1.2. 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. 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. International Initiatives

9.3.1. 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

- 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.
- Vladimir Petrik spent October - Janurary 2020 in Willow as a visiting post-doc within the framework of collaboration with the Intelligent Machine Perception project.
- Mircea Cimpoi spent three weeks in March 2019 in Willow as a visiting post-doc within the framework of collaboration with the Intelligent Machine Perception project.

9.4.1.1. Internships

- Anna Kukleva (Master student, University of Bonn) spent six months in the Willow team working on her Master project under supervision of M. Tapaswi and I. Laptev.

9.4.2. Visits to International Teams

9.4.2.1. Explorer programme

- J.Ponce, multiple visits to CMU's Robotics Institute within the framework of the Gaia associated team

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events: Organisation

10.1.1.1. Member of the Organizing Committees

- I. Laptev was co-organizedr of BMVA Symposium on Video Understanding in London, September 2019.
- P.-L. Guhur was co-organizer of Junior Conference on Data Science and Engineering at Paris-Saclay, September 2019.

10.1.2. Scientific Events: Selection

10.1.2.1. Area chairs

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019 (I. Laptev).
- Neural Information Processing Systems (NeurIPS), 2019 (J. Sivic).

10.1.2.2. Member of the Conference Program Committees / Reviewer

- IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019 (A. Miech, M. Tapaswi).
- IEEE/CVF International Conference on Computer Vision (ICCV), 2019 (A. Miech, M. Tapaswi).
- Neural Information Processing Systems (NeurIPS), 2019 (I. Laptev, I. Rocco, M. Tapaswi).
- British Machine Vision Conference (BMVC), 2019 (I. Rocco).
- IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019 (I. Laptev, J. Carpentier, J. Sivic, Y. Labbe).
- IEEE International Conference on Robotics and Automation (ICRA), 2019 (J. Carpentier, Y. Labbe, J. Sivic).

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 (M. Tapaswi).
- IEEE Transactions on Pattern Analysis and Machine Intelligence (I. Rocco).
- IEEE Transactions on Robotics (J. Carpentier).
- IEEE Robotics and Automation Letters (J. Carpentier).
- Plos One (M. Tapaswi).

10.1.4. Invited Talks

- I. Laptev, invited talk, AI Video Summit, FAIR, Los Angeles, June 2019.
- I. Laptev, invited talk, ActivityNet CVPR Workshop, Long Beach, June 2019.
- I. Laptev, keynote, AIST, Kazan, July 2019.
- I. Laptev, invited talk, ELLIS Workshop, San Sebastian, September 2019.
- I. Laptev, seminar, KTH, Stockholm, September 2019.
- I. Laptev, invited talk, BMVA Symposium on Video Understanding, London, September 2019.
- I. Laptev, invited talk, Qualcomm-UvA Deep Learning Seminars, Amsterdam, September 2019.
- I. Laptev, invited talk, Extreme Vision ICCV Workshop, Seoul, October 2019.
- I. Laptev, invited talk, CoVieW ICCV Workshop, Seoul, October 2019.
- I. Laptev, invited talk, Hands in Action ICCV Workshop, Seoul, October 2019.
- I. Laptev, keynote, DICTA, Perth, December 2019.
- I. Laptev, invited talk, SkolTech, Moscow, December 2019.
- J. Ponce, invited talk, Pavillon des Sciences CCI, Montbelliard, March 2019.
- J. Ponce, invited talk, France is AI, Paris, Oct. 2019.
- J. Ponce, invited talk, Ecole navale, Brest, Jan. 2019.
- J. Sivic, Keynote, European Conference on Mobile Robots (ECMR), Prague, September 2019.
- J. Sivic, Seminar on Weakly supervised learning for visual recognition, College de France, February 2019
- J. Sivic, Invited speaker, ML in PL conference, Warsaw, November 2019
- J. Sivic, Seminar, DeepMind, London, February 2019

- J. Sivic, Talk at invited workshop of the ELLIS computer vision program, San Sebastian, September 2019
- I. Rocco, invited talk, SMILE seminar at Télécom Paris, Paris, March 2019.
- I. Rocco, seminar, IMAGINE group at ENPC, Paris, April 2019.
- J-P. Laumond, invited talk, La robotique, Café Scientifique, Cité des sciences et de l'industrie, Paris, Janurary 2019.
- J-P. Laumond, invited talk, Une machine peut-elle apprendre ?, Les Découvrades, Toulouse, February 2019.
- J-P. Laumond, invited talk, La robotique : retour vers le réel, Théatre Tourski, Marseille, February 2019.
- J-P. Laumond, invited talk, Robots, Exhibition opening, Cité des sciences et de l'industrie, Paris, March 2019.
- J-P. Laumond, scientific talk, Robot Motion, Workshop on Robotics and Art, IEEE ICRA, Montreal, May 2019.
- J-P. Laumond, scientific talk, Wording Robotics, Workshop on Rhetoric and Robotics, IEEE ICRA, Montreal, May 2019.
- J-P. Laumond, scientific talk, Robotique et Biomécanique, Keynote, Congrès de physiologie, Montpellier, June 2019.
- J-P. Laumond, scientific talk, Robotics and AI, Istituto di Robotica e Macchine Intelligenti Grand Opening, Roma, October 2019.
- J. Carpentier, scientific talk, La robotique envisagée comme une science des données, CNRS Grenoble, January 2019.
- J. Carpentier, scientific talk, La robotique envisagée comme une science des données, Inria Grenoble, February 2019.
- J. Carpentier, scientific talk, Analytical Derivatives of Rigid Body Dynamics algorithms, RSS, June 2019.
- J. Carpentier, scientific talk, La robotique : la fabrique du mouvement artificiel, Journées DGDI Inria, Rocquencourt, November 2019.
- M. Tapaswi, scientific talk, Understanding Humans and the Stories they Tell, LIMSI Paris, February 2019.
- M. Tapaswi, scientific talk, Understanding Humans and the Stories they Tell, UPC Barcelona, August 2019.

10.1.5. 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 of advisory board, Computer Vision Foundation (J. Sivic).
- Board Member Deputy, European Laboratory for Learning and Intelligent Systems – ELLIS (J. Sivic)

10.1.6. Scientific Expertise

- J. Ponce, coordinator of the AI theme for the joint French-American Committee on Science and Technology, 2018-.
- I. Laptev, head of scientific board at VisionLabs, 2019-.

10.1.7. 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 développement technologique (CDT), 2012-2018 (J. Sivic).
- Member of the Hiring Committee at Computer Science department, École normale supérieure (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 supérieure, 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: J-P. Laumond and J. Carpentier, "Robotics", M1 MPRI, Ecole normale supérieure and Ecole normale superieure de Cachan, 48h.
- Master: J-P. Laumond, "Robotics", Ecole des Mines de Paris, 4h.
- Master: J. Sivic, three lectures (3 x 1.5h) in the 3D computer vision class of V. Hlavac at Charles University in Prague.
- License: P.L. Guhur, "Developping web applications with React Js", L3, Université Grenoble Alpes, 30h.
- J. Ponce co-organized the PRAIRIE AI Summer School in Paris, 2019.
- Bachelor: J. Ponce, "Inroduction to computer vision" MS level class, NYU Center for Data Science, Fall 2019

10.2.2. Supervision

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

PhD in progress : Pierre-Louis Guhur, “Learning Visual Language Manipulation”, started in Oct 2019, I. Laptev and C. Schmid.

PhD in progress : Aamr El Kazdadi, started in Oct 2019, J. Carpentier and J. Ponce.

PhD in progress : Bruno Lecouat, started in Sept 2019, J. Ponce and J. Mairal (Inria Grenoble).

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 : Minttu Alakuijala, started in Feb 2019, J. Ponce and C. Schmid.

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, “Reconstruction and recognition of hand-object manipulations”, started in Nov 2017, I. Laptev and C. Schmid.

PhD in progress : Ronan Riochet, “Reconstruction and recognition of hand-object manipulations”, started in Oct 2017, E. Dupoux, 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.

PhD defended : Gul Varol, “Deep learning methods for video interpretation”, graduated in May 2019, I. Laptev, C. Schmid.

PhD defended : Julia Peyre, “Learning to reason about scenes from images and language”, graduated in August 2019, C. Schmid, I. Laptev, J. Sivic.

PhD defended : Theophile Dalens, “Learning to analyze and reconstruct architectural scenes”, graduated in September 2019, M. Aubry and J. Sivic.

10.2.3. Juries

- PhD thesis committee:
  - Dmitry Ulyanov, SkolTech, Russia, 2019 (I. Laptev, rapporteur).
  - Judith Bütepage, KTH, Sweden, 2019 (I. Laptev, rapporteur).
  - Alexander Richard, University of Bonn, Germany, 2019 (I. Laptev, rapporteur).
  - Rıza Alp Güler, Université Paris Saclay, France, 2019 (I. Laptev, rapporteur).
  - Tatiana Shpakova, l’École Normale Supérieure, France, 2019 (I. Laptev, examinateur).
  - Konstantin SHMELKOV, Universite Grenoble Alpes, (J. Sivic, examinateur)
  - Sanjeel PAREKH, Télécom Paris, (J. Sivic, rapporteur)
  - Filip RADENOVIC, Czech Technical University (J. Sivic, rapporteur)
  - Nathan PIASCO, Université Bourgogne Franche-Comte (J. Sivic, examinateur)
- HDR committee:
  - Karteek Alahari, HDR defense, Jan. 2019 (J. Ponce)
  - Mathieu Aubry, HDR defense, Aug. 2019 (J. Ponce)

10.3. Popularization

10.3.1. Internal or external Inria responsibilities

J.P. Laumond is the scientific curator of the permanent exhibition at Cité des Sciences et de l’Industrie. The aim of this exhibition is to help the general audience to understand the concepts of robotics. Indeed, the notion of robotics today is packed with many preconceived notions, phobias, and utopias, all fed by literature and a rich film culture. The real challenge of the exhibition is the presentation of authentic working robots that raises awareness of our relationship to these singular machines. How do they work? What are they for? What are their performances today and what will they be tomorrow? The exhibition lays bare the actual capabilities of robots and provides insight into the current issues.

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] T. DALENS. *Learnable factored image representations for visual discovery*, Ecole Normale Superieure de Paris - ENS Paris, September 2019, https://hal.inria.fr/tel-02296150

[2] J. PEYRE. *Learning to detect visual relations*, Ecole Normale Superieure de Paris - ENS Paris, August 2019, https://hal.inria.fr/tel-02332673

[3] G. VAROL. *Learning human body and human action representations from visual data*, Ecole Normale Supérieure (ENS) ; ED 386 : École doctorale de sciences mathématiques de Paris centre, UPMC, May 2019, https://hal.inria.fr/tel-02266593

Articles in International Peer-Reviewed Journals

[4] F. BAILLY, J. CARPENTIER, M. BENALLEGUE, B. WATIER, P. SOUÈRES. *Estimating the Center of Mass and the Angular Momentum Derivative for Legged Locomotion — A recursive approach*, in "IEEE Robotics and Automation Letters", October 2019, vol. 4, n^{o} 4, pp. 4155-4162 [*DOI :* 10.1109/LRA.2019.2931200], https://hal.archives-ouvertes.fr/hal-02058890

[5] T. DALENS, M. AUBRY, J. SIVIC. *Bilinear image translation for temporal analysis of photo collections*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", October 2019 [*DOI :* 10.1109/TPAMI.2019.2950317], https://hal.inria.fr/hal-02452091

[6] P. FERNBACH, S. TONNEAU, O. STASSE, J. CARPENTIER, M. TAÏX. *C-CROC: Continuous and Convex Resolution of Centroidal dynamic trajectories for legged robots in multi-contact scenarios*, in "IEEE Transactions on Robotics", January 2020, pp. 1-16, forthcoming [*DOI :* 10.1109/TRO.2020.2964787], https://hal.laas.fr/hal-01894869

[7] X. S. HU, S. ZAGORUYKO, N. KOMODAKIS. *Exploring Weight Symmetry in Deep Neural Networks*, in "Computer Vision and Image Understanding", August 2019, vol. 187, https://arxiv.org/abs/1812.11027 [*DOI :* 10.1016/J.CVIU.2019.07.006], https://hal.archives-ouvertes.fr/hal-01978633

[8] A. TORII, H. TAIRA, J. SIVIC, M. POLLEFEYS, M. OKUTOMI, T. PAJDLA, T. SATTLER. *Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization?*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", September 2019, vol. XX, pp. 1-14 [*DOI :* 10.1109/TPAMI.2019.2941876], https://hal.inria.fr/hal-02448029

International Conferences with Proceedings

[9] R. BUDHIRAJA, J. CARPENTIER, N. MANSARD. *Dynamics Consensus between Centroidal and Whole-Body Models for Locomotion of Legged Robots*, in "ICRA 2019 - IEEE International Conference on Robotics and Automation", Montreal, Canada, May 2019, https://hal.laas.fr/hal-01875031

[10] 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

[11] *Best Paper*
H. CISNEROS, J. SIVIC, T. MIKOLOV. *Evolving Structures in Complex Systems*, in "SSCI 2019 - IEEE Symposium Series on Computational Intelligence", Xiamen, China, December 2019, https://arxiv.org/abs/1911.01086 - IEEE Symposium Series on Computational Intelligence 2019 (IEEE SSCI 2019), https://hal.inria.fr/hal-02448134.

[12] Y. DING, J. YANG, J. PONCE, H. KONG. *An Efficient Solution to the Homography-Based Relative Pose Problem With a Common Reference Direction*, in "ICCV 2019 - International Conference on Computer Vision", Seoul, South Korea, October 2019, https://hal.archives-ouvertes.fr/hal-02266544

[13] M. DUSMANU, I. ROCCO, T. PAJDLA, M. POLLEFEYS, J. SIVIC, A. TORII, T. SATTLER. *D2-Net: A Trainable CNN for Joint Detection and Description of Local Features*, in "CVPR 2019 - IEEE Conference on Computer Vision and Pattern Recognition", Long Beach, United States, June 2019, https://arxiv.org/abs/1905.03561 - Accepted at CVPR 2019, https://hal.archives-ouvertes.fr/hal-02438461

[14] Y. HASSON, G. VAROL, D. TZIONAS, I. KALEVATYKH, M. J. BLACK, I. LAPTEV, C. SCHMID. *Learning joint reconstruction of hands and manipulated objects*, in "CVPR 2019 - IEEE Conference on Computer Vision and Pattern Recognition", Long Beach, United States, IEEE, June 2019, pp. 1-14, https://hal.archives-ouvertes.fr/hal-02429093

[15] J. LEE, D. KIM, J. S. PONCE, B. HAM. *SFNet: Learning Object-aware Semantic Correspondence*, in "CVPR 2019 - Computer Vision and Pattern Recognition", Longbeach, United States, June 2019, https://hal.archives-ouvertes.fr/hal-02088666

[16] A. MIECH, D. ZHUKOV, J.-B. ALAYRAC, M. TAPASWI, I. LAPTEV, J. SIVIC. *HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips*, in "ICCV 2019 - International Conference on Computer Vision", Séoul, South Korea, International Conference on Computer Vision, October 2019, https://arxiv.org/abs/1906.03327 - Accepted at ICCV 2019, https://hal.archives-ouvertes.fr/hal-02433497

[17] J. MIN, J. LEE, J. PONCE, M. CHO. *Hyperpixel Flow: Semantic Correspondence with Multi-layer Neural Features*, in "ICCV 2019 - International Conference on Computer Vision", Seoul, South Korea, October 2019, https://hal.archives-ouvertes.fr/hal-02267044

[18] A. PASHEVICH, R. STRUDEL, I. KALEVATYKH, I. LAPTEV, C. SCHMID. *Learning to Augment Synthetic Images for Sim2Real Policy Transfer*, in "IROS 2019 - IEEE/RSJ International Conference on Intelligent Robots and Systems", Macao, China, November 2019, pp. 1-6, https://arxiv.org/abs/1903.07740 - 7 pages, https://hal.archives-ouvertes.fr/hal-02273326

[19] J. PEYRE, I. LAPTEV, C. SCHMID, J. SIVIC. *Detecting unseen visual relations using analogies*, in "ICCV 2019 - International Conference on Computer Vision", Seoul, South Korea, October 2019, https://arxiv.org/abs/1812.05736v3 , https://hal.archives-ouvertes.fr/hal-01975760

[20] R. STRUDEL, A. PASHEVICH, I. KALEVATYKH, I. LAPTEV, J. SIVIC, C. SCHMID. *Learning to combine primitive skills: A step towards versatile robotic manipulation*, in "ICRA", Paris, France, May 2020, https://arxiv.org/abs/1908.00722 - 11 pages, https://hal.archives-ouvertes.fr/hal-02274969

[21] H. TAIRA, I. ROCCO, J. SEDLAR, M. OKUTOMI, J. SIVIC, T. PAJDLA, T. SATTLER, A. TORII. *Is This The Right Place? Geometric-Semantic Pose Verification for Indoor Visual Localization*, in "ICCV 2019 - International Conference on Computer Vision", Seoul, South Korea, October 2019, https://arxiv.org/abs/1908.04598 , https://hal.archives-ouvertes.fr/hal-02438468

[22] M. TAPASWI, M. T. LAW, S. FIDLER. *Video Face Clustering with Unknown Number of Clusters*, in "ICCV 2019 - International Conference on Computer Vision", Seoul, South Korea, October 2019, https://arxiv.org/abs/1908.03381 - Accepted to ICCV 2019, code and data at https://github.com/makarandtapaswi/BallClustering_ICCV2019, https://hal.inria.fr/hal-02435407

[23] M. TRAGER, M. HEBERT, J. PONCE. *Coordinate-Free Carlsson-Weinshall Duality and Relative Multi-ViewGeometry*, in "CVPR 2019 - IEEE Conference on Computer Vision and Pattern Recognition", Long Beach, United States, June 2019, https://hal.inria.fr/hal-01676732

[24] Y. ZHANG, Y. GAO, S. GU, Y. GUO, M. LIU, Z. SUN, Z. HOU, H. YANG, Y. WANG, J. YANG, J. PONCE, H. KONG. *Build your own hybrid thermal/EO camera for autonomous vehicle*, in "ICRA 2019 - IEEE International Conference on Robotics and Automation", Montreal, Canada, May 2019, to appear in the Proc. of the IEEE International Conference on Robotics and Automation, 2019, https://hal.inria.fr/hal-02051880

[25] D. ZHUKOV, J.-B. ALAYRAC, R. G. CINBIS, D. F. FOUHEY, I. LAPTEV, J. SIVIC. *Cross-task weakly supervised learning from instructional videos*, in "CVPR 2019 - IEEE Conference on Computer Vision and Pattern Recognition", Long Beach, CA, United States, June 2019, https://hal.archives-ouvertes.fr/hal-02434806

Other Publications

[26] 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

[27] 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

[28] B. KIM, J. PONCE, B. HAM. *Deformable Kernel Networks for Joint Image Filtering*, January 2020, working paper or preprint [*DOI :* 10.1007/s11263-018-1074-6], https://hal.archives-ouvertes.fr/hal-01857016

[29] B. LECOUAT, J. PONCE, J. MAIRAL. *Revisiting Non Local Sparse Models for Image Restoration*, December 2019, working paper or preprint, https://hal.inria.fr/hal-02414291

[30] J. LEE, D. KIM, W. LEE, J. PONCE, B. HAM. *Learning Semantic Correspondence Exploiting an Object-level Prior*, January 2020, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02433226

[31] C. MASTALLI, R. BUDHIRAJA, W. MERKT, G. SAUREL, B. HAMMOUD, M. NAVEAU, J. CARPENTIER, S. VIJAYAKUMAR, N. MANSARD. *Crocoddyl: An Efficient and Versatile Framework for Multi-Contact Optimal Control*, September 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02294059

[32] A. MIECH, J.-B. ALAYRAC, L. SMAIRA, I. LAPTEV, J. SIVIC, A. ZISSERMAN. *End-to-End Learning of Visual Representations from Uncurated Instructional Videos*, January 2020, https://arxiv.org/abs/1912.06430 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02433504

[33] 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

[34] A. MIECH, I. LAPTEV, J. SIVIC, H. WANG, L. TORRESANI, D. TRAN. *Leveraging the Present to Anticipate the Future in Videos*, January 2020, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02433506

[35] G. VAROL, I. LAPTEV, C. SCHMID, A. ZISSERMAN. *Synthetic Humans for Action Recognition from Unseen Viewpoints*, January 2020, https://arxiv.org/abs/1912.04070 - working paper or preprint, https://hal.inria.fr/hal-02435731

[36] 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

[37] S. ZAGORUYKO, Y. LABBÉ, I. KALEVATYKH, I. LAPTEV, J. CARPENTIER, M. AUBRY, J. SIVIC. *Monte-Carlo Tree Search for Efficient Visually Guided Rearrangement Planning*, April 2019, https://arxiv.org/abs/1904.10348 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02108930