IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2017

Project-Team WILLOW

Models of visual object recognition and scene understanding

IN COLLABORATION WITH: Département d'Informatique de l'Ecole Normale Supérieure

RESEARCH CENTER
Paris

THEME
Vision, perception and multimedia interpretation

Table of contents

Project-Team WILLOW

Creation of the Project-Team: 2007 June 01

Keywords:

Computer Science and Digital Science:

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

Other Research Topics and Application Domains:

B9.4.1. - Computer science
B9.4.5. - Data science

1. Personnel

Research Scientists

Ivan Laptev [Inria, Senior Researcher, HDR]
Josef Sivic [Inria, Senior Researcher, HDR]

Faculty Member

Jean Ponce [Team leader, ENS Paris, Professor]

Post-Doctoral Fellow

Anton Osokin [Inria]

PhD Students

Guilhem Cheron [Inria]
Theophile Dalens [Inria]
Thomas Eboli [ENS]
Yana Hasson [Inria]
Vadim Kantorov [Inria]
Zongmian Li [Inria]
Antoine Miech [Inria]
Maxime Oquab [Inria]
Julia Peyre [Inria]
Ronan Riochet [Inria]
Ignacio Rocco Spremolla [Inria]
Rafael Sampaio de Rezende [Inria]
Matthew Trager [Inria]
Gül Varol [Inria]
Tuan-Hung Vu [Inria]
Dmitry Zhukov [Inria]

Technical staff

Igor Kalevatykh [Inria]
Mauricio Diaz [Inria]

Administrative Assistants
Sabrine Boumizy [Inria]
Sarah Le [Inria]

Visiting Scientists
Hildegard Kuehne [Universtiy of Bonn, Apr 2017]
Jason Corso [University of Michigan, Apr 2017]
Alexei Efros [UC Berkeley, June 2017]

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 four new Phd students: Thomas Eboli (ENS), Yana Hasson (Inria), Zongmian Li (Inria) and Dmitry Zhukov (Inria). Alexei Efros (Professor, UC Berkeley, USA) visited Willow during June. Hildegard Kuehne (Universtiy of Bonn) and Jason Corso (University of Michigan) visited Willow during April.

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 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.

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

3.4. Human activity capture and classification

From a scientific point of view, visual action understanding is a computer vision problem that until recently has received little attention outside of extremely specific contexts such as surveillance or sports. Many of the current approaches to the visual interpretation of human activities are designed for a limited range of operating conditions, such as static cameras, fixed scenes, or restricted actions. The objective of this part of our project is to attack the much more challenging problem of understanding actions and interactions in unconstrained video depicting everyday human activities such as in sitcoms, feature films, or news segments. The recent emergence of automated annotation tools for this type of video data (Everingham, Sivic, Zisserman, 2006; Laptev, Marszałek, Schmid, Rozenfeld, 2008; Duchenne, Laptev, Sivic, Bach, Ponce, 2009) means that massive amounts of labelled data for training and recognizing action models will at long last be available.

Our research agenda in this scientific domain is described below and our recent results are outlined in detail in Section 7.4.

- **Weakly-supervised learning and annotation of human actions in video.** We aim to leverage the huge amount of video data using readily-available annotations in the form of video scripts. Scripts, however, often provide only imprecise and incomplete information about the video. We address this problem with weakly-supervised learning techniques both at the text and image levels.
- **Descriptors for video representation.** Video representation has a crucial role for recognizing human actions and other components of a visual scene. Our work in this domain aims to develop generic methods for representing video data based on realistic assumptions. In particular, we develop deep learning methods and design new trainable representations for various tasks such as human action recognition, person detection, segmentation and tracking.

4. Application Domains

4.1. Introduction

We believe that foundational modeling work should be grounded in applications. This includes (but is not restricted to) the following high-impact domains.

4.2. Quantitative image analysis in science and humanities

We plan to apply our 3D object and scene modeling and analysis technology to image-based modeling of human skeletons and artifacts in anthropology, and large-scale site indexing, modeling, and retrieval in archaeology and cultural heritage preservation. Most existing work in this domain concentrates on image-based rendering, that is, the synthesis of good-looking pictures of artifacts and digs. We plan to focus instead on quantitative applications. We are engaged in a project involving the archaeology laboratory at ENS and focusing on image-based artifact modeling and decorative pattern retrieval in Pompeii. Application of our 3D reconstruction technology is now being explored in the field of cultural heritage and archeology by the start-up Iconem, founded by Y. Ubelmann, a Willow collaborator.

4.3. Video Annotation, Interpretation, and Retrieval

Both specific and category-level object and scene recognition can be used to annotate, augment, index, and retrieve video segments in the audiovisual domain. The Video Google system developed by Sivic and Zisserman (2005) for retrieving shots containing specific objects is an early success in that area. A sample application, suggested by discussions with Institut National de l'Audiovisuel (INA) staff, is to match set photographs with actual shots in film and video archives, despite the fact that detailed timetables and/or annotations are typically not available for either medium. Automatically annotating the shots is of course also relevant for archives that may record hundreds of thousands of hours of video. Some of these applications will be pursued in our MSR-Inria project.

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

- J. Sivic (together with J. Philbin, O. Chum, M. Isard, and A. Zisserman) received the Longuet-Higgins Prize for "Fundamental contributions in Computer Vision", awarded at the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
- J. Sivic (together with A. Zisserman) received the Helmholtz Prize for "fundamental contributions to computer vision", awarded at the International Conference on Computer Vision, 2017.
- J. Sivic (together with B. Russell, A. Efros, B. Freeman and A. Zisserman) received the Helmholtz Prize for "fundamental contributions to computer vision", awarded at the International Conference on Computer Vision, 2017.
- I. Laptev (together with T. Lindeberg) received the Helmholtz Prize for "fundamental contributions to computer vision", awarded at the International Conference on Computer Vision, 2017.

6. New Software and Platforms

6.1. LOUPE

Learnable mOdUle for Pooling fEatures
KEYWORDS: Video analysis - Computer vision
FUNCTIONAL DESCRIPTION: LOUPE (Learnable mOdUle for Pooling fEatures) is a Tensorflow toolbox that implements several modules for pooling features such as NetVLAD, NetRVLAD, NetFV and Soft-DBoW. It also allows to use their Gated version. This toolbox was mainly use in the winning approach of the Youtube 8M Large Scale Video Understanding challenge

- Participants: Antoine Miech, Ivan Laptev and Josef Sivic
- Contact: Antoine Miech
- Publication: Learning from Video and Text via Large-Scale Discriminative Clustering
- URL: https://github.com/antoine77340/LOUPE

6.2. object-states-action

KEYWORD: Computer vision
FUNCTIONAL DESCRIPTION: Code for the paper Joint Discovery of Object States and Manipulation Actions, ICCV 2017: Many human activities involve object manipulations aiming to modify the object state. Examples of common state changes include full/empty bottle, open/closed door, and attached/detached car wheel. In this work, we seek to automatically discover the states of objects and the associated manipulation actions. Given a set of videos for a particular task, we propose a joint model that learns to identify object states and to localize state-modifying actions. Our model is formulated as a discriminative clustering cost with constraints. We assume a consistent temporal order for the changes in object states and manipulation actions, and introduce new optimization techniques to learn model parameters without additional supervision. We demonstrate successful discovery of seven manipulation actions and corresponding object states on a new dataset of videos depicting real-life object manipulations. We show that our joint formulation results in an improvement of object state discovery by action recognition and vice versa.

- Participants: Jean-Baptiste Alayrac, Josef Sivic, Ivan Laptev and Simon Lacoste-Julien
- Contact: Jean-Baptiste Alayrac
- Publication: Joint Discovery of Object States and Manipulation Actions
- URL: https://github.com/jalayrac/object-states-action

6.3. SURREAL

Learning from Synthetic Humans
KEYWORDS: Synthetic human - Segmentation - Neural networks
FUNCTIONAL DESCRIPTION: The SURREAL dataset consisting of synthetic videos of humans, and models trained on this dataset are released in this package. The code for rendering synthetic images of people and for training models is also included in the release.

- Participants: Gül Varol Simsekli, Xavier Martin, Ivan Laptev and Cordelia Schmid
- Contact: Gül Varol Simsekli
- Publication: Learning from Synthetic Humans
- URL: http://www.di.ens.fr/willow/research/surreal/

6.4. UNREL

Weakly-supervised learning of visual relations
KEYWORDS: Recognition - Computer vision
FUNCTIONAL DESCRIPTION: Open source release of the software package for the ICCV17 paper by Peyre et al. "Weakly-supervised learning of visual relations". The package provides a full implementation of the method (training and evaluation) and the release of the UnRel dataset. Links to all of these are available at the project page http://www.di.ens.fr/willow/research/unrel/

- Participants: Julia Peyre, Ivan Laptev, Cordelia Schmid and Josef Sivic
- Contact: Julia Peyre
- Publication: Weakly-supervised learning of visual relations
- URL: http://www.di.ens.fr/willow/research/unrel/

6.5. BIOGAN

GANs for Biological Image Synthesis
KEYWORDS: Computer vision - Biology
FUNCTIONAL DESCRIPTION: This software package implements the method in the ICCV 2017 paper by Osokin et al. "GANs for Biological Image Synthesis".

- Participants: Federico Vaggi, Anton Osokin and Anatole Chessel
- Contact: Anton Osokin
- Publication: GANs for Biological Image Synthesis

6.6. KernelImageRetrieval

Kernel square-loss exemplar machines for image retrieval
KEYWORD: Computer vision
FUNCTIONAL DESCRIPTION: This software package contains the code for the CVPR'17 paper by Rezende et al. "Kernel square-loss exemplar machines for image retrieval". It provides the implementation of all variants of the pipeline as well as the trained parameters for each of the tested base features.

- Participants: Jean Ponce, Francis Bach, Patrick Pérez and Rafael Sampaio De Rezende
- Contact: Rafael Sampaio De Rezende
- Publication: Kernel Square-Loss Exemplar Machines for Image Retrieval
- URL: https://github.com/rafarez/slem/

6.7. SCNet

SCNet: Learning semantic correspondence

KEYWORD: Computer vision
FUNCTIONAL DESCRIPTION: This software package implements the method for the ICCV'17 paper by Han et al. "SCNet: Learning Semantic Correspondence". The package provides the code, the training and testing subsets and the trainable architecture.

- Participants: Rafael Sampaio De Rezende, Bumsub Ham, Minsu Cho, Cordelia Schmid and Jean Ponce
- Contact: Rafael Sampaio De Rezende
- Publication: SCNet: Learning Semantic Correspondence
- URL: https://github.com/k-han/SCNet/

6.8. CNNGeometric

Convolutional neural network architecture for geometric matching
KEYWORD: Computer vision
FUNCTIONAL DESCRIPTION: Open source release of the software package for the CVPR'17 paper by Rocco et al. "Convolutional neural network architecture for geometric matching". This release provides a full implementation of the method, including code for training models, and testing on standard datasets, as well as trained models.

- Participants: Ignacio Rocco Spremolla, Relja Arandjelovic and Josef Sivic
- Contact: Ignacio Rocco Spremolla
- Publication: Convolutional neural network architecture for geometric matching
- URL: http://www.di.ens.fr/willow/research/cnngeometric/

6.9. LSDClustering

Large-Scale Discriminative Clustering
KEYWORDS: Video analysis - Computer vision
FUNCTIONAL DESCRIPTION: This software package implements the method in the ICCV'17 paper by Miech et al. "Learning from Video and Text via Large-Scale Discriminative Clustering".

- Participants: Antoine Miech, Jean-Baptiste Alayrac, Piotr Bojanowski, Ivan Laptev and Josef Sivic
- Contact: Antoine Miech
- Publication: Learning from Video and Text via Large-Scale Discriminative Clustering
- URL: http://www.di.ens.fr/willow/research/learningvideotext/

7. New Results

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

7.1.1. Congruences and Concurrent Lines in Multi-View Geometry

Participants: Jean Ponce, Bernd Sturmfels, Matthew Trager.

We present a new framework for multi-view geometry in computer vision. A camera is a mapping between $\mathbb{P}^3$ and a line congruence. This model, which ignores image planes and measurements, is a natural abstraction of traditional pinhole cameras. It includes two-slit cameras, pushbroom cameras, catadioptric cameras, and many more (Figure 1). We study the concurrent lines variety, which consists of n-tuples of lines in $\mathbb{P}^3$ that intersect at a point. Combining its equations with those of various congruences, we derive constraints for corresponding images in multiple views. We also study photographic cameras which use image measurements and are modeled as rational maps from $\mathbb{P}^3$ to $\mathbb{P}^2$ or $\mathbb{P}^1 \times \mathbb{P}^1$. This work has been published in [7].

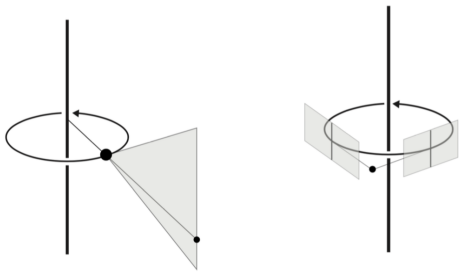

Figure 1. Non-central panoramic (left) and stereo panoramic cameras (right) are examples of non-linear cameras that can be modeled using line congruences.

7.1.2. General models for rational cameras and the case of two-slit projections

Participants: Matthew Trager, Bernd Sturmfels, John Canny, Martial Hebert, Jean Ponce.

The rational camera model provides a general methodology for studying abstract nonlinear imaging systems and their multi-view geometry. This paper builds on this framework to study "physical realizations" of rational cameras. More precisely, we give an explicit account of the mapping between between physical visual rays and image points, which allows us to give simple analytical expressions for direct and inverse projections (Figure 2). We also consider "primitive" camera models, that are orbits under the action of various projective transformations, and lead to a general notion of intrinsic parameters. The methodology is general, but it is illustrated concretely by an in-depth study of two-slit cameras, that we model using pairs of linear projections. This simple analytical form allows us to describe models for the corresponding primitive cameras, to introduce intrinsic parameters with a clear geometric meaning, and to define an epipolar tensor characterizing two-view correspondences. In turn, this leads to new algorithms for structure from motion and self-calibration. This work has been published in [22].

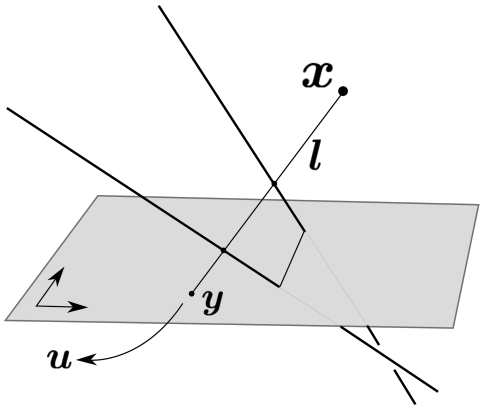

Figure 2. A general camera associates a scene point x with a visual ray l, then maps the ray l to its intersection y with some retinal plane π, and finally uses a projective coordinate system on π to express y as a point u in $\mathbb{P}^2$.

7.1.3. Changing Views on Curves and Surfaces

Participants: Kathlén Kohn, Bernd Sturmfels, Matthew Trager.

In this paper, visual events in computer vision are studied from the perspective of algebraic geometry. Given a sufficiently general curve or surface in 3-space, we consider the image or contour curve that arises by projecting from a viewpoint. Qualitative changes in that curve occur when the viewpoint crosses the visual event surface (Figure 3). We examine the components of this ruled surface, and observe that these coincide with the iterated singular loci of the coisotropic hypersurfaces associated with the original curve or surface. We derive formulas, due to Salmon and Petitjean, for the degrees of these surfaces, and show how to compute exact representations for all visual event surfaces using algebraic methods. This work was published in [6].

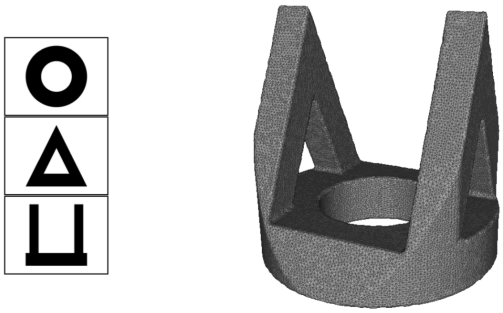

Figure 3. Changing views of a curve correspond to Reidemeister moves. The viewpoint z crosses the tangential surface (left), edge surface (middle), or trisecant surface (right).

7.1.4. On point configurations, Carlsson-Weinshall duality, and multi-view geometry

Participants: Matthew Trager, Martial Hebert, Jean Ponce.

We propose in this project projective point configurations as a natural setting for studying perspective projection in a geometric, coordinate-free manner. We show that classical results on the effect of permutations on point configurations give a purely synthetic formulation of the well known analytical Carlsson-Weinshall duality between camera pinholes and scene points. We further show that the natural parameterizations of configurations in terms of subsets of their points provides a new and simple analytical formulation of Carlsson-Weinshall duality in any scene and image coordinate systems, not just in the reduced coordinate frames used traditionally. When working in such reduced coordinate systems, we give a new and complete characterization of multi-view geometry in terms of a reduced joint image and its dual. We also introduce a new parametrization of trinocular geometry in terms of reduced trilinearities, and show that, unlike trifocal tensors, these are not subject to any nonlinear internal constraints. This leads to purely linear primal and dual structure-from-motion algorithms, that we demonstrate with a preliminary implementation on real data. This work has been submitted to CVPR'18 [27].

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

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

Accurate visual localization is a key technology for autonomous navigation. 3D structure-based methods employ 3D models of the scene to estimate the full 6DOF 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 this paper, 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 pose accuracy similar to the state-of-the-art structure-based methods. Our results, published at [21] and illustrated in Figure 4, suggest that we might want to reconsider the current approach for accurate large-scale localization.

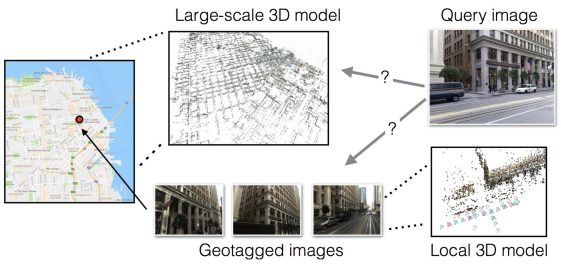

Figure 4. ***Large-scale visual localization.*** *2D image-based methods (bottom) use image retrieval and return the pose of the most relevant database image. 3D structure-based methods (top) use 2D-3D matches against a 3D model for camera pose estimation. Both approaches have been developed largely independently of each other and never compared properly before. We provide such comparison in this work.*

7.2. Category-level object and scene recognition

7.2.1. SCNet: Learning semantic correspondence

Participants: Kai Han, Rafael S. Rezende, Bumsub Ham, Kwan-Yee K. Wong, Minsu Cho, Cordelia Schmid, Jean Ponce.

In this work we propose a convolutional neural network architecture, called SCNet, for learning a geometrically plausible model for establishing semantic correspondence between images depicting different instances of the same object or scene category. SCNet uses region proposals as matching primitives, and explicitly incorporates geometric consistency in its loss function. An overview of the architecture can be seen in Figure 5. It is trained on image pairs obtained from the PASCAL VOC 2007 keypoint dataset, and a comparative evaluation on several standard benchmarks demonstrates that the proposed approach substantially outperforms both recent deep learning architectures and previous methods based on hand-crafted features. This work has been published in [13].

7.2.2. Kernel square-loss exemplar machines for image retrieval

Participants: Rafael S. Rezende, Joaquin Zepeda, Jean Ponce, Francis Bach, Patrick Pérez.

In this work we explore the promise of an exemplar classifier, such as exemplar SVM (ESVM), as a feature encoder for image retrieval and extends this approach in several directions: We first show that replacing the hinge loss by the square loss in the ESVM cost function significantly reduces encoding time with negligible effect on accuracy. We call this model square-loss exemplar machine, or SLEM. An overview of the pipeline can be seen in Figure 6. We then introduce a kernelized SLEM which can be implemented efficiently through low-rank matrix decomposition, and displays improved performance. Both SLEM variants exploit the fact that the negative examples are fixed, so most of the SLEM computational complexity is relegated to an offline process independent of the positive examples. Our experiments establish the performance and computational advantages of our approach using a large array of base features and standard image retrieval datasets. This work has been published in [19].

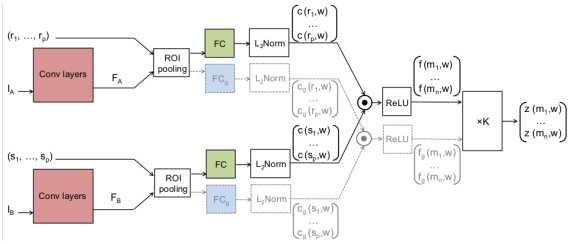

Figure 5. The SCNet architectures. Three variants are proposed: SCNet-AG, SCNet-A, and SCNet-AG+. The basic architecture, SCNet-AG, is drawn in solid lines. Colored boxes represent layers with learning parameters and the boxes with the same color share the same parameters. "$\times K$" denotes the voting layer for geometric scoring. A simplified variant, SCNet-A, learns appearance information only by making the voting layer an identity function. An extended variant, SCNet-AG+, contains an additional stream drawn in dashed lines. SCNet-AG learns a single embedding c for both appearance and geometry, whereas SCNet-AG+ learns an additional and separate embedding c_g for geometry.

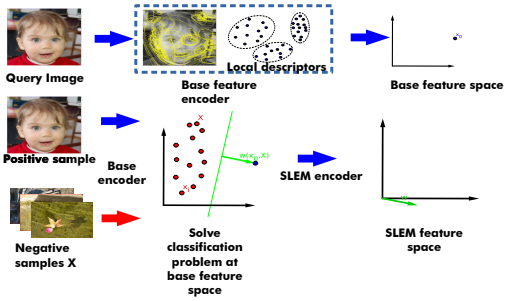

Figure 6. Pipeline of SLEM. First row encapsulates the construction of a base feature for a query image, which usually consists of extracting, embedding and aggregating local descriptors into a vector, here written as x_0. After repeating the process of base feature calculation a database of sample images and obtaining a matrix X of base features, we solve a exemplar classifier by labeling x_0 as the lonely positive example (called exemplar) and the columns of X as negatives. The solution ω to this classification problem, which is a function of x_0 and X, is our SLEM encoding of the query image.

7.2.3. Weakly-supervised learning of visual relations

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

This paper introduces a novel approach for modeling visual relations between pairs of objects. We call relation a triplet of the form $(subject, predicate, object)$ where the predicate is typically a preposition (eg. 'under', 'in front of') or a verb ('hold', 'ride') that links a pair of objects $(subject, object)$. Learning such relations is challenging as the objects have different spatial configurations and appearances depending on the relation in which they occur. Another major challenge comes from the difficulty to get annotations, especially at box-level, for all possible triplets, which makes both learning and evaluation difficult. The contributions of this paper are threefold. First, we design strong yet flexible visual features that encode the appearance and spatial configuration for pairs of objects. Second, we propose a weakly-supervised discriminative clustering model to learn relations from image-level labels only. Third we introduce a new challenging dataset of unusual relations (UnRel) together with an exhaustive annotation, that enables accurate evaluation of visual relation retrieval. We show experimentally that our model results in state-of-the-art results on the visual relationship dataset significantly improving performance on previously unseen relations (zero-shot learning), and confirm this observation on our newly introduced UnRel dataset. This work has been published in [18] and example results are shown in Figure 7.

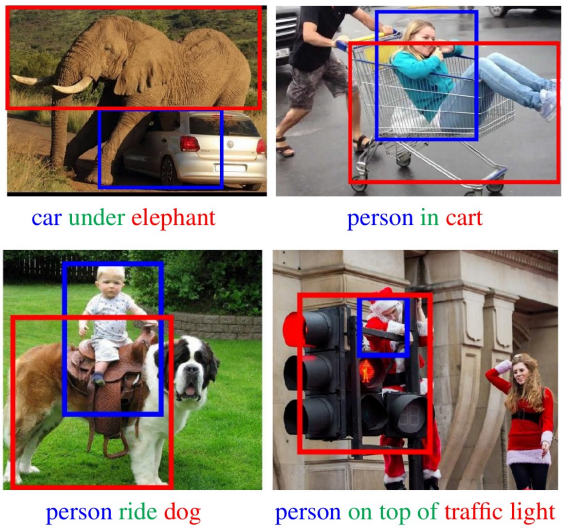

Figure 7. Examples of top retrieved pairs of boxes in UnRel dataset for unusual queries with our weakly supervised model

7.2.4. Convolutional neural network architecture for geometric matching

Participants: Ignacio Rocco, Relja Arandjelović, Josef Sivic.

We address the problem of determining correspondences between two images in agreement with a geometric model such as an affine or thin-plate spline transformation, and estimating its parameters. The contributions of this work are three-fold. First, we propose a convolutional neural network architecture for geometric matching, illustrated in Figure 8. The architecture is based on three main components that mimic the standard steps of feature extraction, matching and simultaneous inlier detection and model parameter estimation, while being trainable end-to-end. Second, we demonstrate that the network parameters can be trained from synthetically generated imagery without the need for manual annotation and that our matching layer significantly increases

generalization capabilities to never seen before images. Finally, we show that the same model can perform both instance-level and category-level matching giving state-of-the-art results on the challenging Proposal Flow dataset. This work has been published in [20].

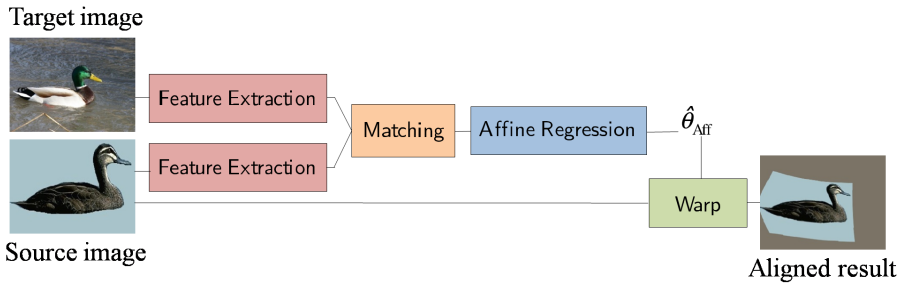

Figure 8. Proposed CNN architecture for geometric matching. Source and target images are passed through feature extraction networks which have tied parameters, followed by a matching network which matches the descriptors. The output of the matching network is passed through a regression network which outputs the parameters of the geometric transformation, which are used to produce the final alignment.

7.3. Image restoration, manipulation and enhancement

7.3.1. GANs for Biological Image Synthesis

Participants: Anton Osokin, Anatole Chessel, Rafael E. Carazo Salas, Federico Vaggi.

In this work we propose a novel application of Generative Adversarial Networks (GAN) to the synthesis of cells imaged by fluorescence microscopy. Compared to natural images, cells tend to have a simpler and more geometric global structure that facilitates image generation. However, the correlation between the spatial pattern of different fluorescent proteins reflects important biological functions, and synthesized images have to capture these relationships to be relevant for biological applications. We adapt GANs to the task at hand and propose new models with casual dependencies between image channels that can generate multi-channel images, which would be impossible to obtain experimentally (see Figure 9). We evaluate our approach using two independent techniques and compare it against sensible baselines. Finally, we demonstrate that by interpolating across the latent space we can mimic the known changes in protein localization that occur through time during the cell cycle, allowing us to predict temporal evolution from static images. This paper has been published in [17].

7.4. Human activity capture and classification

7.4.1. Learning from Synthetic Humans

Participants: Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael Black, Ivan Laptev, Cordelia Schmid.

Estimating human pose, shape, and motion from images and video are fundamental challenges with many applications. Recent advances in 2D human pose estimation use large amounts of manually-labeled training data for learning convolutional neural networks (CNNs). Such data is time consuming to acquire and difficult to extend. Moreover, manual labeling of 3D pose, depth and motion is impractical. In [23], we present SURREAL: a new large-scale dataset with synthetically-generated but realistic images of people rendered from 3D sequences of human motion capture data. We generate more than 6 million frames together with ground truth pose, depth maps, and segmentation masks. We show that CNNs trained on our synthetic dataset

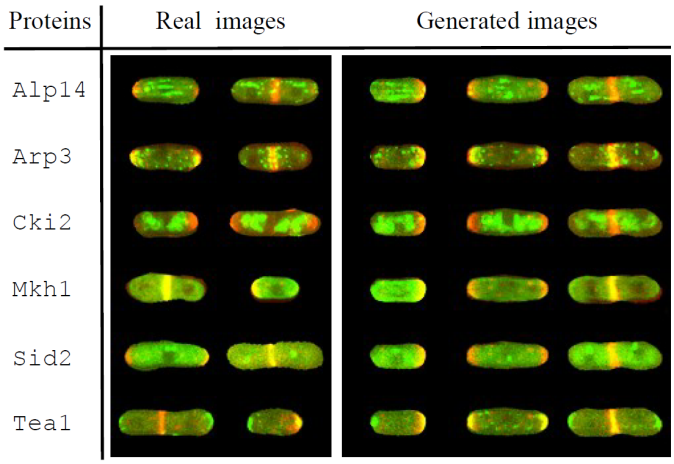

Figure 9. Real (left) and generated (right) images of fission yeast cells with protein bgs4 depicted in the red channel and 6 other proteins depicted in the green channel. The synthetic images were generated with our star-shaped GAN. The star-shaped model can generate multiple green channels aligned with the same red channel whereas the training images have only one green channel.

allow for accurate human depth estimation and human part segmentation in real RGB images, see Figure 10. Our results and the new dataset open up new possibilities for advancing person analysis using cheap and large-scale synthetic data. This work has been published in [23].

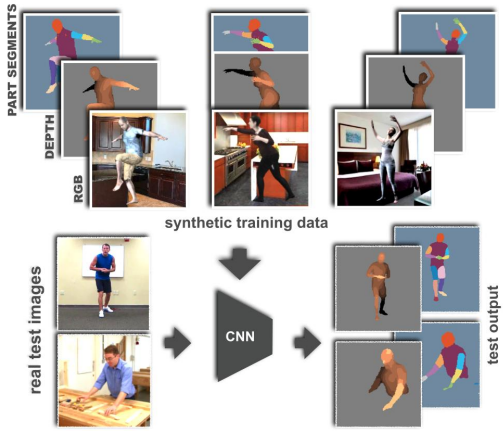

Figure 10. We generate photo-realistic synthetic images and their corresponding ground truth for learning pixel-wise classification problems: human parts segmentation and depth estimation. The convolutional neural network trained only on synthetic data generalizes on real images sufficiently for both tasks.

7.4.2. Learning from Video and Text via Large-Scale Discriminative Clustering

Participants: Miech Antoine, Alayrac Jean-Baptiste, Bojanowski Piotr, Laptev Ivan, Sivic Josef.

Discriminative clustering has been successfully applied to a number of weakly-supervised learning tasks. Such applications include person and action recognition, text-to-video alignment, object co-segmentation and colocalization in videos and images. One drawback of discriminative clustering, however, is its limited scalability. We address this issue and propose an online optimization algorithm based on the Block-Coordinate Frank-Wolfe algorithm. We apply the proposed method to the problem of weakly supervised learning of actions and actors from movies together with corresponding movie scripts. The scaling up of the learning problem to 66 feature length movies enables us to significantly improve weakly supervised action recognition. Figure 11 illustrates output of our method on movies. This work has been published in [15]

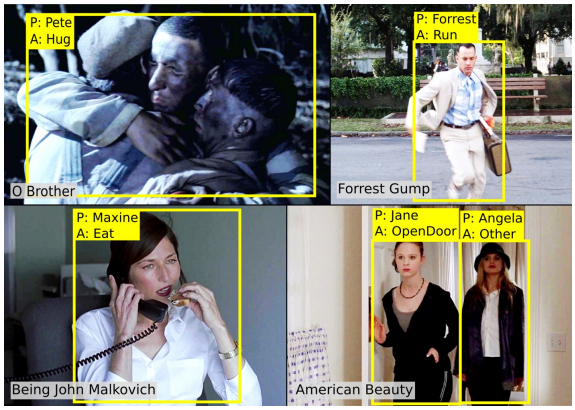

Figure 11. We automatically recognize actors and their actions in a of dataset of 66 movies with scripts as weak supervision

7.4.3. ActionVLAD: Learning spatio-temporal aggregation for action classification

Participants: Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, Bryan Russell.

In this work, we introduce a new video representation for action classification that aggregates local convolutional features across the entire spatio-temporal extent of the video. We do so by integrating state-of-the-art two-stream networks [42] with learnable spatio-temporal feature aggregation [6]. The resulting architecture is end-to-end trainable for whole-video classification. We investigate different strategies for pooling across space and time and combining signals from the different streams. We find that: (i) it is important to pool jointly across space and time, but (ii) appearance and motion streams are best aggregated into their own separate representations. Finally, we show that our representation outperforms the two-stream base architecture by a large margin (13out-performs other baselines with comparable base architec-tures on HMDB51, UCF101, and Charades video classification benchmarks. The work has been published at [12] and the method is illustrated in Figure 12.

7.4.4. Localizing Moments in Video with Natural Language

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

We consider retrieving a specific temporal segment, or moment, from a video given a natural language text description. Methods designed to retrieve whole video clips with natural language determine what occurs in a video but not when. To address this issue, we propose the Moment Context Network (MCN) which effectively localizes natural language queries in videos by integrating local and global video features over time. A key obstacle to training our MCN model is that current video datasets do not include pairs of localized video segments and referring expressions, or text descriptions which uniquely identify a corresponding moment. Therefore, we collect the Distinct Describable Moments (DiDeMo) dataset which consists of over 10,000

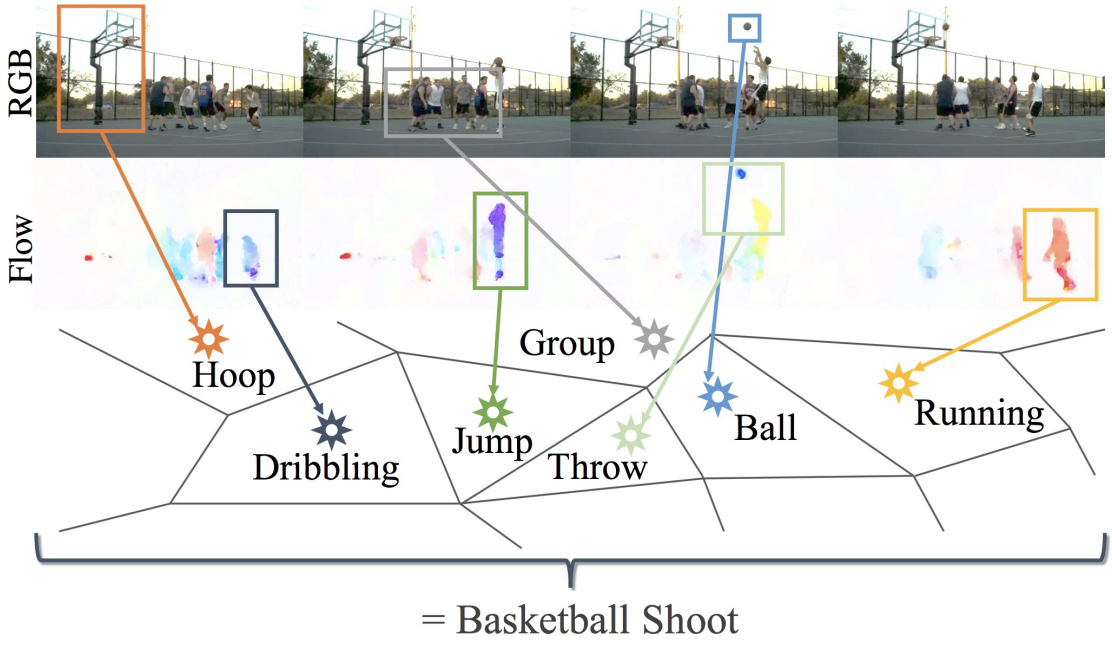

Figure 12. How do we represent actions in a video? We propose ActionVLAD, a spatio-temporal aggregation of a set of action primitives over the appearance and motion streams of a video. For example, a basketball shoot may be represented as an aggregation of appearance features corresponding to 'group of players', 'ball' and 'basketball hoop'; and motion features corresponding to 'run', 'jump', and 'shoot'.

unedited, personal videos in diverse visual settings with pairs of localized video segments and referring expressions. We demonstrate that MCN outperforms several baseline methods and believe that our initial results together with the release of DiDeMo will inspire further research on localizing video moments with natural language. The work has been published at [14] and results are illustrated in Figure 13.

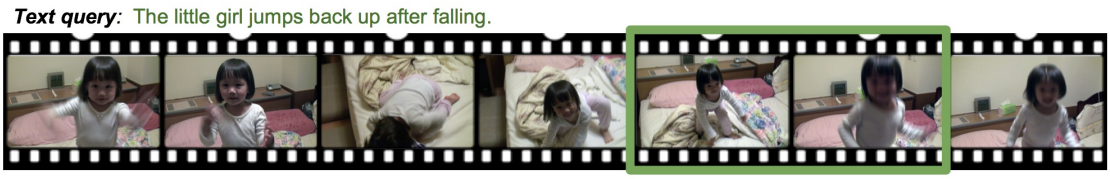

Figure 13. We consider localizing moments in video with natural language and demonstrate that incorporating local and global video features is important for this task. To train and evaluate our model, we collect the Distinct Describable Moments (DiDeMo) dataset which consists of over 40,000 pairs of localized video moments and corresponding natural language.

7.4.5. Learnable pooling with Context Gating for video classification

Participants: Miech Antoine, Laptev Ivan, Sivic Josef.

Common video representations often deploy an average or maximum pooling of pre-extracted frame features over time. Such an approach provides a simple means to encode feature distributions, but is likely to be suboptimal. As an alternative, in this work we explore combinations of learnable pooling techniques such as Soft Bag-of-words, Fisher Vectors, NetVLAD, GRU and LSTM to aggregate video features over time. We also introduce a learnable non-linear network unit, named Context Gating, aiming at modeling interdependencies between features. The overview of our network architecture is illustrated in Figure 14. We evaluate the method on the multi-modal Youtube-8M Large-Scale Video Understanding dataset using pre-extracted visual and audio features. We demonstrate improvements provided by the Context Gating as well as by the combination of learnable pooling methods. We finally show how this leads to the best performance, out of more than 600 teams, in the Kaggle Youtube-8M Large-Scale Video Understanding challenge. This work has been published in [26].

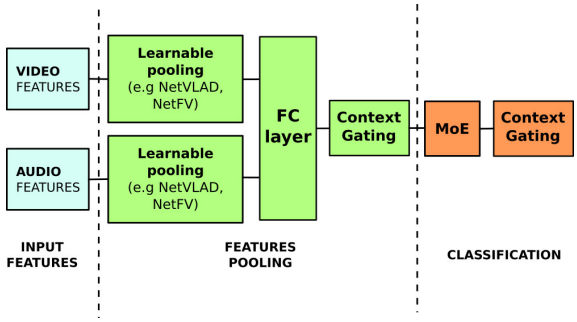

Figure 14. Overview of our network architecture for video classification

8. Bilateral Contracts and Grants with Industry

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

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

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

8.2. Google: Learning to annotate videos from movie scripts (Inria)

Participants: Josef Sivic, Ivan Laptev, Jean Ponce.

The goal of this project is to automatically generate annotations of complex dynamic events in video. We wish to deal with events involving multiple people interacting with each other, objects and the scene, for example people at a party in a house. The goal is to generate structured annotations going beyond simple text tags. Examples include entire text sentences describing the video content as well as bounding boxes or segmentations spatially and temporally localizing the described objects and people in video. This is an extremely challenging task due to large intra-class variation of human actions. We propose to learn joint video and text representations enabling such annotation capabilities from feature length movies with coarsely aligned shooting scripts. Building on our previous work in this area, we aim to develop structured representations of video and associated text enabling to reason both spatially and temporally about scenes, objects and people as well as their interactions. Automatic understanding and interpretation of video content is a key-enabling factor for a range of practical applications such as content-aware advertising or search. Novel video and text representations are needed to enable breakthrough in this area.

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

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

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

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

Participants: Guilhem Cheron, Ivan Laptev, Maxime Oquab, Jean Ponce, Josef Sivic, Cordelia Schmid [Inria Lear].

This collaborative project brings together the WILLOW and LEAR project-teams with MSR researchers in Cambridge and elsewhere. The concept builds on several ideas articulated in the “2020 Science” report, including the importance of data mining and machine learning in computational science. Rather than focusing only on natural sciences, however, we propose here to expand the breadth of e-science to include humanities and social sciences. The project we propose will focus on fundamental computer science research in computer vision and machine learning, and its application to archaeology, cultural heritage preservation, environmental science, and sociology, and it will be validated by collaborations with researchers and practitioners in these fields.

In October 2013 a new agreement has been signed for 2013-2017 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.

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. Agence Nationale de la Recherche (ANR): SEMAPOLIS

Participants: Mathieu Aubry, Josef Sivic.

The goal of the SEMAPOLIS project is to develop advanced large-scale image analysis and learning techniques to semantize city images and produce semantized 3D reconstructions of urban environments, including proper rendering. Geometric 3D models of existing cities have a wide range of applications, such as navigation in virtual environments and realistic sceneries for video games and movies. A number of players (Google, Microsoft, Apple) have started to produce such data. However, the models feature only plain surfaces, textured from available pictures. This limits their use in urban studies and in the construction industry, excluding in practice applications to diagnosis and simulation. Besides, geometry and texturing are often wrong when there are invisible or discontinuous parts, e.g., with occluding foreground objects such as trees, cars or lampposts, which are pervasive in urban scenes. This project will go beyond the plain geometric models by producing semantized 3D models, i.e., models which are not bare surfaces but which identify architectural elements such as windows, walls, roofs, doors, etc. Semantic information is useful in a larger number of scenarios, including diagnosis and simulation for building renovation projects, accurate shadow impact taking into account actual window location, and more general urban planning and studies such as solar cell deployment. Another line of applications concerns improved virtual cities for navigation, with object-specific rendering, e.g., specular surfaces for windows. Models can also be made more compact, encoding object repetition (e.g., windows) rather than instances and replacing actual textures with more generic ones according to semantics; it allows cheap and fast transmission over low- bandwidth mobile phone networks, and efficient storage in GPS navigation devices.

This is a collaborative effort with LIGM / ENPC (R. Marlet), University of Caen (F. Jurie), Inria Sophia Antipolis (G. Drettakis) and Acute3D (R. Keriven).

9.2. European Initiatives

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

Participant: Ivan Laptev.

WILLOW will be funded in part from 2013 to 2017 by the ERC Starting Grant "Activia" awarded to Ivan Laptev by the European Research Council.

‘Computer vision is concerned with the automated interpretation of images and video streams. Today’s research is (mostly) aimed at answering queries such as ’Is this a picture of a dog?’, (classification) or sometimes ’Find the dog in this photo’ (detection). While categorisation and detection are useful for many tasks, inferring correct class labels is not the final answer to visual recognition. The categories and locations of objects do not provide direct understanding of their function i.e., how things work, what they can be used for, or how they can act and react. Such an understanding, however, would be highly desirable to answer currently unsolvable queries such as ’Am I in danger?’ or ’What can happen in this scene?’. Solving such queries is the aim of this proposal. My goal is to uncover the functional properties of objects and the purpose of actions by addressing visual recognition from a different and yet unexplored perspective. The main novelty of this proposal is to leverage observations of people, i.e., their actions and interactions to automatically learn the use, the purpose and the function of objects and scenes from visual data. The project is timely as it builds upon the two key recent technological advances: (a) the immense progress in visual recognition of objects, scenes and human actions achieved in the last ten years, as well as (b) the emergence of a massive amount of public image and video data now available to train visual models. ACTIVIA addresses fundamental research issues in automated interpretation of dynamic visual scenes, but its results are expected to serve as a basis for ground-breaking technological advances in practical applications. The recognition of functional properties and intentions as explored in this project will directly support high-impact applications such as detection of abnormal events, which are likely to revolutionise today’s approaches to crime protection, hazard prevention, elderly care, and many others.’

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

Participant: Josef Sivic.

The contract has begun on Nov 1st 2014. WILLOW will be funded in part from 2014 to 2018 by the ERC Starting Grant "Leap" awarded to Josef Sivic by the European Research Council.

‘People constantly draw on past visual experiences to anticipate future events and better understand, navigate, and interact with their environment, for example, when seeing an angry dog or a quickly approaching car. Currently there is no artificial system with a similar level of visual analysis and prediction capabilities. LEAP is a first step in that direction, leveraging the emerging collective visual memory formed by the unprecedented amount of visual data available in public archives, on the Internet and from surveillance or personal cameras - a complex evolving net of dynamic scenes, distributed across many different data sources, and equipped with plentiful but noisy and incomplete metadata. The goal of this project is to analyze dynamic patterns in this shared visual experience in order (i) to find and quantify their trends; and (ii) learn to predict future events in dynamic scenes. With ever expanding computational resources and this extraordinary data, the main scientific challenge is now to invent new and powerful models adapted to its scale and its spatio-temporal, distributed and dynamic nature. To address this challenge, we will first design new models that generalize across different data sources, where scenes are captured under vastly different imaging conditions such as camera viewpoint, temporal sampling, illumination or resolution. Next, we will develop a framework for finding, describing and quantifying trends that involve measuring long-term changes in many related scenes. Finally, we will develop a methodology and tools for synthesizing complex future predictions from aligned past visual experiences. Our models will be automatically learnt from large-scale, distributed, and asynchronous visual data, coming from different sources and with different forms of readily-available but noisy and incomplete metadata such as text, speech, geotags, scene depth (stereo sensors), or gaze and body motion (wearable sensors). Breakthrough progress on these problems would have profound implications on our everyday lives as well as science and commerce, with safer cars that anticipate the behavior of pedestrians on streets; tools that help doctors monitor, diagnose and predict patients’ health; and smart glasses that help people react in unfamiliar situations enabled by the advances from this project.’

9.3. International Initiatives

9.3.1. IMPACT: Intelligent machine perception

Participants: Josef Sivic, Jean Ponce, Ivan Laptev.

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

9.3.2. Inria CityLab initiative

Participants: Josef Sivic, Jean Ponce, Ivan Laptev, Alexei Efros [UC Berkeley].

Willow participates in the ongoing CityLab@Inria initiative (co-ordinated by V. Issarny), which aims to leverage Inria research results towards developing “smart cities" by enabling radically new ways of living in, regulating, operating and managing cities. The activity of Willow focuses on urban-scale quantitative visual analysis and is pursued in collaboration with A. Efros (UC Berkeley).

Currently, map-based street-level imagery, such as Google Street-view provides a comprehensive visual record of many cities worldwide. Additional visual sensors are likely to be wide-spread in near future: cameras will be built in most manufactured cars and (some) people will continuously capture their daily visual experience using wearable mobile devices such as Google Glass. All this data will provide large-scale, comprehensive and dynamically updated visual record of urban environments.

The goal of this project is to develop automatic data analytic tools for large-scale quantitative analysis of such dynamic visual data. The aim is to provide quantitative answers to questions like: What are the typical architectural elements (e.g., different types of windows or balconies) characterizing a visual style of a city district? What is their geo-spatial distribution? How does the visual style of a geo-spatial area evolve over time? What are the boundaries between visually coherent areas in a city? Other types of interesting questions concern distribution of people and their activities: How do the number of people and their activities at particular places evolve during a day, over different seasons or years? Are there tourists sightseeing, urban dwellers shopping, elderly walking dogs, or children playing on the street? What are the major causes for bicycle accidents?

Break-through progress on these goals would open-up completely new ways smart cities are visualized, modeled, planned and simulated, taking into account large-scale dynamic visual input from a range of visual sensors (e.g., cameras on cars, visual data from citizens, or static surveillance cameras).

9.3.3. 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

Prof. Alexei Efros (UC Berkeley, USA) visited Willow during June. Hildegard Kuehne (Universtiy of Bonn) and Jason Corso (University of Michigan) visited Willow during April.

9.4.1.1. Internships

Kai Han has visited Willow from the University of Hong Kong.

9.4.2. Visits to International Teams

9.4.2.1. Research Stays Abroad

Jean Ponce is visiting New York University since September 2017.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. General Chair, Scientific Chair

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

10.1.1.2. Member of the Organizing Committees

- M. Trager is an organizer of “Minisymposium” on “Algebraic Vision” at the SIAM conference on Applied Algebraic Geometry (Atlanta, July 31 – August 4, 2017).
- G. Varol is an organizer of “Women in Computer Vision Workshop” at IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
- G. Varol is an organizer of “Multiview Relationships in 3D Data Workshop” at International Conference on Computer Vision (ICCV), 2017.

10.1.2. Scientific Events Selection

10.1.2.1. Area chairs

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018 (J. Sivic).
- International Conference on Computer Vision (ICCV), 2017 (J. Sivic).
- European Conference on Computer Vision (ECCV), 2018 (I. Laptev, J. Sivic).

10.1.2.2. Member of the Conference Program Committees

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 (G. Cheron, I. Laptev, A. Osokin, J. Sivic).
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018 (J.-B. Alayrac, M. Oquab, R. Rezende, I. Rocco, G. Varol).
- International Conference on Computer Vision (ICCV), 2017 (I. Laptev, A. Osokin).
- Neural Information Processing Systems (NIPS), 2017 (A. Osokin, J. Sivic).
- International Conference on Learning Representations (ICLR), 2017 (J. Sivic).
- International Conference on Machine Learning (ICML), 2017 (A. Osokin).
- IEEE International Conference on Robotics and Automation (ICRA), 2018 (A. Miech).

10.1.3. Journals

10.1.3.1. Member of the editorial board

- International Journal of Computer Vision (I. Laptev, J. Ponce, J. Sivic).
- IEEE Transactions on Pattern Analysis and Machine Intelligence (I. Laptev, J. Sivic).
- Foundations and Trends in Computer Graphics and Vision (J. Ponce).
- I. Laptev co-edit a special issue on "Deep Learning for Computer Vision" in Computer Vision and Image Understanding.

10.1.3.2. Reviewer

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

10.1.4. Others

- J. Sivic is senior fellow of the Neural Computation and Adaptive Perception program of the Canadian Institute of Advanced Research.
- A. Bursuc obtained the outstanding reviewer award at International Conference on Computer Vision (ICCV), 2017.

10.1.5. Invited Talks

- I. Laptev, Seminar, Inria Rennes, December, 2017.
- I. Laptev, Invited talk, Chalearn Workshop on Action, Gesture, and Emotion Recognition, Venice, October, 2017.
- I. Laptev, Invited talk, The Joint Video and Language Understanding Workshop, Venice, October, 2017.
- I. Laptev, Invited talk, Rentrée ENS, Paris, September, 2017
- I. Laptev, Invited talk, ML Day, Pré-GDR IA, Paris, September 2017.
- I. Laptev, Invited talk, Frontiers of Video Technology workshop, Adobe, July, 2017.
- I. Laptev, Invited talk, Workshop on YouTube-8M Large-Scale Video Understanding, Honolulu, July, 2017.
- I. Laptev, Invited talk, Workshop on Visual Understanding Across Modalities, Honolulu, July, 2017.
- I. Laptev, Plenary talk, Iberian Conference on Pattern Recognition and Image Analysis, Faro, June, 2017.
- I. Laptev, Invited talk, Paris ML Meetup Spatio-temporal Series Hackathon, Paris, February, 2017.
- A. Miech, Invited Talk, Facebook AI Research, Paris, September 2017.
- A. Miech, Invited Talk, DGA TIM2017 Seminar, Paris, July 2017.

- A. Miech, Invited Talk, Paris ML Meetup, Paris, June 2017.
- J. Ponce, Keynote speaker, Korean Conference on Computer Vision, Seoul, June 2017.
- J. Ponce, Invited talk, BioVision, Lyon, April 2017.
- J. Ponce, Invited talk, Dept. of computer science of the University of Central Florida, Orlando, February 2017.
- J. Ponce, Invited talk, New York University, Center for Data Science, September 2017.
- J. Ponce, Invited talk, Facebook AI Research, New York, October 2017.
- J. Ponce, Invited talk, Amazon, Seattle, November 2017.
- J. Ponce, Invited talk, AEF conference, Paris, 2017.
- J. Ponce, Invited talk, BIOVISION, The World Life Sciences Forum, Lyon, April 2017.
- J. Ponce, Invited talk, DGSI, 2017.
- J. Ponce, Invited talk, DRM 20th Anniversary, 2017.
- J. Sivic, Seminar, KTH Stockholm, January, 2017.
- J. Sivic, Invited talk, ParisTech Telecom, January, 2017.
- J. Sivic, Invited talk, University of Amsterdam, March, 2017.
- J. Sivic, Invited talk, ORASIS, journées francophones des jeunes chercheurs en vision par ordinateur, June, 2017.
- J. Sivic, Invited talk, Visual Understanding for Interaction workshop, CVPR 2017, July, 2017.
- J. Sivic, Invited talk, Frontiers of Video Technology workshop, Adobe, July, 2017.
- J. Sivic, Invited talk, Inria Rennes, September, 2017.
- J. Sivic, Seminar, UC Berkeley, December, 2017.
- J. Sivic, Invited talk, the CIFAR workshop, Long Beach, December 2017.

10.1.6. Leadership within the Scientific Community

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

10.1.7. Scientific Expertise

- J. Sivic gave an overview of state-of-the-art in computer vision at the seminar on deep learning at Academie des Technologies, Paris, November 2017.

10.1.8. Research Administration

- Member, Bureau du comité des projets, Inria, Paris (J. Ponce)
- Director, Department of Computer Science, Ecole normale supérieure (J. Ponce)
- Member, Scientific academic council, PSL Research University (J. Ponce)
- Member, Research representative committee, PSL Research University (J. Ponce).
- Member of Inria Commission de developpement technologique (CDT), 2012- (J. Sivic).
- Member of the Hiring Committe for the tenure track position at CentraleSupelec (I. Laptev).
- Member of the Hiring Committee for Professor of Computer Vision at CentraleSupelec (I. Laptev).

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

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

10.2.2. Supervision

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

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

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

PhD in progress : Dmitry Zhukov, "Learning from instruction videos for personal assistants", started in Oct 2017, I. Laptev and J. Sivic.

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

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

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

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

PhD in progress : Jean-Baptiste Alayrac, "Structured learning from video and natural language", started in 2014, I. Laptev, J. Sivic and S. Lacoste-Julien (Inria SIERRA / U. Montreal).

PhD : Rafael Sampaio de Rezende, "New methods for image classification, image retrieval and semantic correspondence", graduated in 2017, J. Ponce.

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

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

PhD in progress : Vadim Kantorov, "Large-scale video mining and recognition", started in 2012, I. Laptev.

PhD in progress : Maxime Oquab, "Learning to annotate dynamic scenes with convolutional neural networks", started in Jan 2014, L. Bottou (Facebook AI Research), I. Laptev and J. Sivic.

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

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

10.2.3. Juries

PhD thesis committee:

- Ahmet Iscen, University of Rennes, France, 2017, (J. Sivic, rapporteur).
- Edouard Oyallon, ENS, France, 2017, (I. Laptev, examinatuer).
- Juan Manuel PÉREZ RÚA, Inria Rennes, France, 2017, (I. Laptev, examinatuer).
- Ali Razavian, KTH Stockholm, Sweden, 2017 (J. Sivic, reviewer).
- Francisco Suzano Massa, ENPC, France, 2017, (J. Sivic, examinatuer).
- Mattis Paulin, Universite de Grenoble, France, 2017 (J. Sivic, rapporteur).

10.3. Popularization

- Interview with Science et Vie, June 2017 (J. Ponce).
- Interview with Forbes Russia, June 2017 (J. Ponce).
- Interview "Science et Vie Junior", July 2017 (J. Ponce).
- Interview for "Le jaune et le rouge", Oct. 2017 (J. Ponce).
- Interview for the "les 100 français de l'IA" dossier of Usine Nouvelle, Dec. 2017 (J. Ponce).
- Two articles in "Binaires" for the newspaper Le Monde (J. Ponce).

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] R. SAMPAIO DE REZENDE. *New methods for image classification, image retrieval and semantic correspondence* , École normale supérieure de Paris, December 2017, https://hal.inria.fr/tel-01676893

Articles in International Peer-Reviewed Journals

[2] J.-B. ALAYRAC, P. BOJANOWSKI, N. AGRAWAL, J. SIVIC, I. LAPTEV, S. LACOSTE-JULIEN. *Learning from narrated instruction videos*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", September 2017, vol. XX, https://hal.archives-ouvertes.fr/hal-01580630

[3] R. ARANDJELOVIĆ, P. GRONAT, A. TORII, T. PAJDLA, J. SIVIC. *NetVLAD: CNN architecture for weakly supervised place recognition*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", June 2017, vol. XX, https://hal.inria.fr/hal-01557234

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

[5] B. HAM, M. CHO, C. SCHMID, J. PONCE. *Proposal Flow: Semantic Correspondences from Object Proposals*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2018, forthcoming, https://hal.inria.fr/hal-01644132

[6] K. KOHN, B. STURMFELS, M. TRAGER. *Changing Views on Curves and Surfaces*, in "Acta Mathematica Vietnamica", December 2017, https://arxiv.org/abs/1707.01877 - 31 pages [*DOI :* 10.1007/s40306-017-0240-1], https://hal.inria.fr/hal-01676208

[7] J. PONCE, B. STURMFELS, M. TRAGER. *Congruences and Concurrent Lines in Multi-View Geometry*, in "Advances in Applied Mathematics", 2017, vol. 88, pp. 62-91, https://arxiv.org/abs/1608.05924v2 , https://hal.inria.fr/hal-01423057

[8] A. TORII, R. ARANDJELOVIC, J. SIVIC, M. OKUTOMI, T. PAJDLA. *24/7 place recognition by view synthesis*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", February 2017, 14 p. [*DOI :* 10.1109/TPAMI.2017.2667665], https://hal.inria.fr/hal-01616660

[9] G. VAROL, I. LAPTEV, C. SCHMID. *Long-term Temporal Convolutions for Action Recognition*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2017, https://arxiv.org/abs/1604.04494 , forthcoming [*DOI :* 10.1109/TPAMI.2017.2712608], https://hal.inria.fr/hal-01241518

[10] Y. ZHANG, Y. SU, J. YANG, J. PONCE, H. KONG. *When Dijkstra meets vanishing point: a stereo vision approach for road detection*, in "IEEE Transactions on Image Processing", 2018, pp. 1-12, forthcoming, https://hal.archives-ouvertes.fr/hal-01678548

International Conferences with Proceedings

[11] J.-B. ALAYRAC, J. SIVIC, I. LAPTEV, S. LACOSTE-JULIEN. *Joint Discovery of Object States and Manipulation Actions*, in "ICCV 2017 - IEEE International Conference on Computer Vision", Venice, Italy, October 2017, https://arxiv.org/abs/1702.02738 - Appears in: International Conference on Computer Vision 2017 (ICCV 2017). 15 pages, https://hal.archives-ouvertes.fr/hal-01676084

[12] R. GIRDHAR, D. RAMANAN, A. GUPTA, J. SIVIC, B. RUSSELL. *ActionVLAD: Learning spatio-temporal aggregation for action classification*, in "IEEE Conference on Computer Vision and Pattern Recognition", Honolulu, United States, 2017, https://arxiv.org/abs/1704.02895 - Project page: https://rohitgirdhar.github.io/ActionVLAD/, https://hal.inria.fr/hal-01678686

[13] K. K. HAN, R. S. REZENDE, B. HAM, K.-Y. K. WONG, M. CHO, C. S. SCHMID, J. S. PONCE. *SCNet: Learning Semantic Correspondence*, in "International Conference on Computer Vision", Venise, Italy, International conference on computer vision, October 2017, https://arxiv.org/abs/1705.04043 , https://hal.archives-ouvertes.fr/hal-01576117

[14] L. A. HENDRICKS, O. WANG, E. SHECHTMAN, J. SIVIC, T. DARRELL, B. RUSSELL. *Localizing Moments in Video with Natural Language*, in "IEEE International Conference on Computer Vision - ICCV 2017", Venice, Italy, October 2017, https://arxiv.org/abs/1708.01641 , https://hal.inria.fr/hal-01678699

[15] A. MIECH, J.-B. ALAYRAC, P. BOJANOWSKI, I. LAPTEV, J. SIVIC. *Learning from Video and Text via Large-Scale Discriminative Clustering*, in "ICCV 2017 - IEEE International Conference on Computer Vision", Venice, Italy, October 2017, https://arxiv.org/abs/1707.09074 , https://hal.inria.fr/hal-01569540

[16] A. OSOKIN, F. BACH, S. LACOSTE-JULIEN. *On Structured Prediction Theory with Calibrated Convex Surrogate Losses*, in "The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)", Long Beach, United States, December 2017, https://arxiv.org/abs/1703.02403 , https://hal.archives-ouvertes.fr/hal-01611691

[17] A. OSOKIN, A. CHESSEL, R. E. C. SALAS, F. VAGGI. *GANs for Biological Image Synthesis*, in "ICCV 2017 - IEEE International Conference on Computer Vision", Venice, Italy, October 2017, https://arxiv.org/abs/1708.04692 , https://hal.archives-ouvertes.fr/hal-01611692

[18] J. PEYRE, I. LAPTEV, C. SCHMID, J. SIVIC. *Weakly-supervised learning of visual relations*, in "ICCV 2017-International Conference on Computer Vision 2017", Venice, Italy, October 2017, https://arxiv.org/abs/1707.09472 , https://hal.archives-ouvertes.fr/hal-01576035

[19] R. S. REZENDE, J. ZEPEDA, J. S. PONCE, F. S. BACH, P. PÉREZ. *Kernel Square-Loss Exemplar Machines for Image Retrieval*, in "Computer Vision and Pattern Recognition 2017", Honolulu, United States, Computer vision and pattern recognition 2017, July 2017, https://hal.inria.fr/hal-01515224

[20] I. ROCCO, R. ARANDJELOVIĆ, J. SIVIC. *Convolutional neural network architecture for geometric matching*, in "CVPR 2017 - IEEE Conference on Computer Vision and Pattern Recognition", Honolulu, United States, July 2017, https://arxiv.org/abs/1703.05593 , https://hal.inria.fr/hal-01513001

[21] T. SATTLER, A. TORII, J. SIVIC, M. POLLEFEYS, H. TAIRA, M. OKUTOMI, T. PAJDLA. *Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization?*, in "CVPR 2017 - IEEE Conference on Computer Vision and Pattern Recognition", Honolulu, United States, July 2017, 10 p. , https://hal.inria.fr/hal-01513083

[22] M. TRAGER, B. STURMFELS, J. CANNY, M. HEBERT, J. PONCE. *General models for rational cameras and the case of two-slit projections*, in "CVPR 2017 - IEEE Conference on Computer Vision and Pattern Recognition", Honolulu, United States, July 2017, https://arxiv.org/abs/1612.01160v4 , https://hal.archives-ouvertes.fr/hal-01506996

[23] G. VAROL, J. J. ROMERO, X. MARTIN, N. MAHMOOD, M. J. BLACK, I. LAPTEV, C. SCHMID. *Learning from Synthetic Humans*, in "2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)", Honolulu, United States, July 2017, https://arxiv.org/abs/1701.01370 [*DOI :* 10.1109/CVPR.2017.492], https://hal.inria.fr/hal-01505711

Scientific Books (or Scientific Book chapters)

[24] PARTHENOS (editor). *Digital 3D Objects in Art and Humanities: challenges of creation, interoperability and preservation. White paper: A result of the PARTHENOS Workshop held in Bordeaux at Maison des Sciences de l'Homme d'Aquitaine and at Archeovision Lab. (France), November 30th - December 2nd, 2016*, PARTHENOS, Bordeaux, France, May 2017, 71 p. , https://hal.inria.fr/hal-01526713

Other Publications

[25] R. LEBLOND, J.-B. ALAYRAC, A. OSOKIN, S. LACOSTE-JULIEN. *SEARNN: Training RNNs with global-local losses*, December 2017, https://arxiv.org/abs/1706.04499 - 12 pages, https://hal.inria.fr/hal-01665263

[26] A. MIECH, I. LAPTEV, J. SIVIC. *Learnable pooling with Context Gating for video classification*, June 2017, https://arxiv.org/abs/1706.06905 - working paper or preprint, https://hal.inria.fr/hal-01547378

[27] M. TRAGER, M. HEBERT, J. PONCE. *On point configurations, Carlsson-Weinshall duality, and multi-view geometry*, January 2018, working paper or preprint, https://hal.inria.fr/hal-01676732