Activity Report 2017

Project-Team THOTH

Learning visual models from large-scale data

IN COLLABORATION WITH: Laboratoire Jean Kuntzmann (LJK)

RESEARCH CENTER
Grenoble - Rhône-Alpes

THEME
Vision, perception and multimedia interpretation

Table of contents

Project-Team THOTH

Creation of the Team: 2016 January 01, updated into Project-Team: 2016 March 01

Keywords:

Computer Science and Digital Science:

A3.4. - Machine learning and statistics
A5.3. - Image processing and analysis
A5.4. - Computer vision
A5.9. - Signal processing
A6.2.6. - Optimization
A8.2. - Optimization
A9.2. - Machine learning
A9.3. - Signal analysis
A9.7. - AI algorithmics

Other Research Topics and Application Domains:

B5.6. - Robotic systems
B8.4. - Security and personal assistance
B8.5. - Smart society
B9.4.1. - Computer science
B9.4.5. - Data science

1. Personnel

Research Scientists

Cordelia Schmid [Team leader, Inria, Senior Researcher, HDR]
Karteek Alahari [Inria, Researcher]
Julien Mairal [Inria, Researcher, “en détachement du corps des mines”, HDR]
Grégory Rogez [Inria, Starting Research Position]
Jakob Verbeek [Inria, Senior Researcher, HDR]

Post-Doctoral Fellow

Henrique Morimitsu [Inria]

PhD Students

Alberto Bietti [Univ. Grenoble Alpes, funded by MSR-Inria joint centre, from Sep 2016 to Sep 2019]
Dexiong Chen [Univ. Grenoble-Alpes, from Sep 2017 to Sep 2020]
Nicolas Chesneau [Univ. Grenoble-Alpes, funded by ERC Allegro]
Thomas Dias-Alves [Univ. Grenoble-Alpes, from Sep 2014 until Sep 2017]
Mikita Dvornik [Univ. Grenoble-Alpes, funded by ERC Allegro and ANR Macaron, from Sep 2016 to Sep 2019]
Maha Elbayad [Univ Grenoble Alpes, funded by Persyval DeCoRe project, from Oct 2016 until Sep 2019]
Roman Klokov [Inria, from Sep 2017 until Aug 2020]
Andrei Kulunchakov [Inria, funded by ERC Solaris, from Sep 2017 to Sep 2020]
Hongzhou Lin [Univ. Grenoble-Alpes, from Sep 2014 to Dec 2017]
Pauline Luc [Univ. Grenoble, funded by Facebook, from Jan 2016 to Dec 2018]
Thomas Lucas [Univ. Grenoble Alpes, from Feb 2016 to Sep 2019]
Alexander Pashevich [Inria, from Sep 2017]

Alexandre Sablayrolles [Univ. Grenoble, funded by Facebook, from Mar 2017]
Konstantin Shmelkov [Inria]
Vladyslav Sydorov [Inria]
Pavel Tokmakov [Inria]
Nitika Verma [Univ. Grenoble Alpes, from Sep 2017 to Sep 2020]
Daan Wynen [Inria, from Sep 2015 to Sep 2017]

Technical staff

Ghislain Durif [Inria, from Feb 2017]
Vasiliki Kalogeiton [Inria, until Nov 2017]
Xavier Martin [Inria]

Interns

Dexiong Chen [Inria, from Apr 2017 until Aug 2017]
Vasileios Choutas [Inria, from Apr 2017]
Pau de Jorge Aranda [Inria, from Feb 2017 until Jun 2017]
Andrei Kulunchakov [Inria, from Feb 2017 until Jun 2017]
Erwan Le Roux [Inria, from Feb 2017 until Dec 2017]
Alexander Pashevich [Inria, from Feb 2017 until Jun 2017]
Sergey Rubtsovenko [Inria, from Feb 2017 until Aug 2017]

Administrative Assistant

Nathalie Gillot [Inria]

Visiting Scientists

Francisco Manuel Castro [University of Malaga, Spain, from Mar 2017 until Jun 2017]
Courtney Paquette [University of Washington, Apr 2017]
Juha Ylionas [Aalto University, from Jan 2017 until Mar 2017]
Gunnar Atli Sigurdsson [Carnegie Mellon University, USA, from Jul 2017 until Nov 2017]

External Collaborators

Marco Pedersoli [École de technologie supérieure, Montreal, Canada]
Danila Potapov [until Jan 2017]

2. Overall Objectives

2.1. Overall Objectives

In 2018, it is expected that nearly 80% of the Internet traffic will be due to videos, and that it would take an individual over 5 million years to watch the amount of video that will cross global IP networks each month by then. Thus, there is a pressing and in fact increasing demand to annotate and index this visual content for home and professional users alike. The available text and speech-transcript metadata is typically not sufficient by itself for answering most queries, and visual data must come into play. On the other hand, it is not imaginable to learn the models of visual content required to answer these queries by manually and precisely annotating every relevant concept, object, scene, or action category in a representative sample of everyday conditions—if only because it may be difficult, or even impossible to decide a priori what are the relevant categories and the proper granularity level. This suggests reverting back to the original metadata as source of annotation, despite the fact that the information it provides is typically sparse (e.g., the location and overall topic of newscasts in a video archive) and noisy (e.g., a movie script may tell us that two persons kiss in some scene, but not when, and the kiss may occur off screen or not have survived the final cut). On the other hand, this weak form of "embedded annotation" is rich and diverse, and mining the corresponding visual data from the web, TV or film archives guarantees that it is representative of the many different scene settings depicted in situations typical of on-line content. Thus, leveraging this largely untapped source of information, rather than attempting to hand label all possibly relevant visual data, is a key to the future use of on-line imagery.

Today's object recognition and scene understanding technology operates in a very different setting; it mostly relies on fully supervised classification engines, and visual models are essentially (piecewise) rigid templates learned from hand labeled images. The sheer scale of on-line data and the nature of the embedded annotation call for a departure from this fully supervised scenario. The main idea of the Thoth project-team is to develop a new framework for learning the structure and parameters of visual models by actively exploring large digital image and video sources (off-line archives as well as growing on-line content, with millions of images and thousands of hours of video), and exploiting the weak supervisory signal provided by the accompanying metadata. This huge volume of visual training data will allow us to learn complex non-linear models with a large number of parameters, such as deep convolutional networks and higher-order graphical models. This is an ambitious goal, given the sheer volume and intrinsic variability of the visual data available on-line, and the lack of a universally accepted formalism for modeling it. Yet, the potential payoff is a breakthrough in visual object recognition and scene understanding capabilities. Further, recent advances at a smaller scale suggest that this is realistic. For example, it is already possible to determine the identity of multiple people from news images and their captions, or to learn human action models from video scripts. There has also been recent progress in adapting supervised machine learning technology to large-scale settings, where the training data is very large and potentially infinite, and some of it may not be labeled. Methods that adapt the structure of visual models to the data are also emerging, and the growing computational power and storage capacity of modern computers are enabling factors that should of course not be neglected.

One of the main objective of Thoth is to transform massive visual data into trustworthy knowledge libraries. For that, it addresses several challenges.

- designing and learning structured models capable of representing complex visual information.
- learning visual models from minimal supervision or unstructured meta-data.
- large-scale learning and optimization.

3. Research Program

3.1. Designing and learning structured models

The task of understanding image and video content has been interpreted in several ways over the past few decades, namely image classification, detecting objects in a scene, recognizing objects and their spatial extents in an image, estimating human poses, recovering scene geometry, recognizing activities performed by humans. However, addressing all these problems individually provides us with a partial understanding of the scene at best, leaving much of the visual data unexplained.

One of the main goals of this research axis is to go beyond the initial attempts that consider only a subset of tasks jointly, by developing novel models for a more complete understanding of scenes to address all the component tasks. We propose to incorporate the structure in image and video data explicitly into the models. In other words, our models aim to satisfy the complex sets of constraints that exist in natural images and videos. Examples of such constraints include: (i) relations between objects, like signs for shops indicate the presence of buildings, people on a road are usually walking or standing, (ii) higher-level semantic relations involving the type of scene, geographic location, and the plausible actions as a global constraint, e.g., an image taken at a swimming pool is unlikely to contain cars, (iii) relating objects occluded in some of the video frames to content in other frames, where they are more clearly visible as the camera or the object itself move, with the use of long-term trajectories and video object proposals.

This research axis will focus on three topics. The first is developing deep features for video. This involves designing rich features available in the form of long-range temporal interactions among pixels in a video sequence to learn a representation that is truly spatio-temporal in nature. The focus of the second topic is the challenging problem of modeling human activities in video, starting from human activity descriptors to building intermediate spatio-temporal representations of videos, and then learning the interactions among humans, objects and scenes temporally. The last topic is aimed at learning models that capture the relationships

among several objects and regions in a single image scene, and additionally, among scenes in the case of an image collection or a video. The main scientific challenges in this topic stem from learning the structure of the probabilistic graphical model as well as the parameters of the cost functions quantifying the relationships among its entities. In the following we will present work related to all these three topics and then elaborate on our research directions.

- **Deep features for vision.** Deep learning models provide a rich representation of complex objects but in return have a large number of parameters. Thus, to work well on difficult tasks, a large amount of data is required. In this context, video presents several advantages: objects are observed from a large range of viewpoints, motion information allows the extraction of moving objects and parts, and objects can be differentiated by their motion patterns. We initially plan to develop deep features for videos that incorporate temporal information at multiple scales. We then plan to further exploit the rich content in video by incorporating additional cues, such as the detection of people and their body-joint locations in video, minimal prior knowledge of the object of interest, with the goal of learning a representation that is more appropriate for video understanding. In other words, a representation that is learned from video data and targeted at specific applications. For the application of recognizing human activities, this involves learning deep features for humans and their body-parts with all their spatiotemporal variations, either directly from raw video data or "pre-processed" videos containing human detections. For the application of object tracking, this task amounts to learning object-specific deep representations, further exploiting the limited annotation provided to identify the object.
- **Modeling human activities in videos.** Humans and their activities are not only one of the most frequent and interesting subjects in videos but also one of the hardest to analyze owing to the complexity of the human form, clothing and movements. As part of this task, the Thoth project-team plans to build on state-of-the-art approaches for spatio-temporal representation of videos. This will involve using the dominant motion in the scene as well as the local motion of individual parts undergoing a rigid motion. Such motion information also helps in reasoning occlusion relationships among people and objects, and the state of the object. This novel spatio-temporal representation ultimately provides the equivalent of object proposals for videos, and is an important component for learning algorithms using minimal supervision. To take this representation even further, we aim to integrate the proposals and the occlusion relationships with methods for estimating human pose in videos, thus leveraging the interplay among body-joint locations, objects in the scene, and the activity being performed. For example, the locations of shoulder, elbow and wrist of a person drinking coffee are constrained to move in a certain way, which is completely different from the movement observed when a person is typing. In essence, this step will model human activities by dynamics in terms of both low-level movements of body-joint locations and global high-level motion in the scene.
- **Structured models.** The interactions among various elements in a scene, such as, the objects and regions in it, the motion of object parts or entire objects themselves, form a key element for understanding image or video content. These rich cues define the structure of visual data and how it evolves spatio-temporally. We plan to develop a novel graphical model to exploit this structure. The main components in this graphical model are spatio-temporal regions (in the case of video or simply image regions), which can represent object parts or entire objects themselves, and the interactions among several entities. The dependencies among the scene entities are defined with a higher order or a global cost function. A higher order constraint is a generalization of the pairwise interaction term, and is a cost function involving more than two components in the scene, e.g., several regions, whereas a global constraint imposes a cost term over the entire image or video, e.g., a prior on the number of people expected in the scene. The constraints we plan to include generalize several existing methods, which are limited to pairwise interactions or a small restrictive set of higher-order costs. In addition to learning the parameters of these novel functions, we will focus on learning the structure of the graph itself—a challenging problem that is seldom addressed in current approaches. This provides an elegant way to go beyond state-of-the-art deep learning methods, which are limited to learning the high-level interaction among parts of an object, by learning the relationships among objects.

3.2. Learning of visual models from minimal supervision

Today's approaches to visual recognition learn models for a limited and fixed set of visual categories with fully supervised classification techniques. This paradigm has been adopted in the early 2000's, and within it enormous progress has been made over the last decade.

The scale and diversity in today's large and growing image and video collections (such as, e.g., broadcast archives, and personal image/video collections) call for a departure from the current paradigm. This is the case because to answer queries about such data, it is unfeasible to learn the models of visual content by manually and precisely annotating every relevant concept, object, scene, or action category in a representative sample of everyday conditions. For one, it will be difficult, or even impossible to decide a-priori what are the relevant categories and the proper granularity level. Moreover, the cost of such annotations would be prohibitive in most application scenarios. One of the main goals of the Thoth project-team is to develop a new framework for learning visual recognition models by actively exploring large digital image and video sources (off-line archives as well as growing on-line content), and exploiting the weak supervisory signal provided by the accompanying metadata (such as captions, keywords, tags, subtitles, or scripts) and audio signal (from which we can for example extract speech transcripts, or exploit speaker recognition models).

Textual metadata has traditionally been used to index and search for visual content. The information in metadata is, however, typically sparse (e.g., the location and overall topic of newscasts in a video archive [1]) and noisy (e.g., a movie script may tell us that two persons kiss in some scene, but not when, and the kiss may occur off screen or not have survived the final cut). For this reason, metadata search should be complemented by visual content based search, where visual recognition models are used to localize content of interest that is not mentioned in the metadata, to increase the usability and value of image/video archives. *The key insight that we build on in this research axis is that while the metadata for a single image or video is too sparse and noisy to rely on for search, the metadata associated with large video and image databases collectively provide an extremely versatile source of information to learn visual recognition models.* This form of "embedded annotation" is rich, diverse and abundantly available. Mining these correspondences from the web, TV and film archives, and online consumer generated content sites such as Flickr, Facebook, or YouTube, guarantees that the learned models are representative for many different situations, unlike models learned from manually collected fully supervised training data sets which are often biased.

The approach we propose to address the limitations of the fully supervised learning paradigm aligns with "Big Data" approaches developed in other areas: we rely on the orders-of-magnitude-larger training sets that have recently become available with metadata to compensate for less explicit forms of supervision. This will form a sustainable approach to learn visual recognition models for a much larger set of categories with little or no manual intervention. Reducing and ultimately removing the dependency on manual annotations will dramatically reduce the cost of learning visual recognition models. This in turn will allow such models to be used in many more applications, and enable new applications based on visual recognition beyond a fixed set of categories, such as natural language based querying for visual content. This is an ambitious goal, given the sheer volume and intrinsic variability of the every day visual content available on-line, and the lack of a universally accepted formalism for modeling it. Yet, the potential payoff is a breakthrough in visual object recognition and scene understanding capabilities.

This research axis is organized into the following three sub-tasks:

- **Weakly supervised learning.** For object localization we will go beyond current methods that learn one category model at a time and develop methods that learn models for different categories concurrently. This allows "explaining away" effects to be leveraged, i.e., if a certain region in an image has been identified as an instance of one category, it cannot be an instance of another category at the same time. For weakly supervised detection in video we will consider detection proposal methods. While these are effective for still images, recent approaches for the spatio-temporal domain need further improvements to be similarly effective. Furthermore, we will exploit appearance and

[1] For example at the Dutch national broadcast archive Netherlands Institute of Sound and Vision, with whom we collaborated in the EU FP7 project AXES, typically one or two sentences are used in the metadata to describe a one hour long TV program.

motion information jointly over a set of videos. In the video domain we will also continue to work on learning recognition models from subtitle and script information. The basis of leveraging the script data which does not have a temporal alignment with the video, is to use matches in the narrative in the script and the subtitles (which do have a temporal alignment with the video). We will go beyond simple correspondences between names and verbs relating to self-motion, and match more complex sentences related to interaction with objects and other people. To deal with the limited amount of occurrences of such actions in a single movie, we will consider approaches that learn action models across a collection of movies.

- **Online learning of visual models.** As a larger number of visual category models is being learned, online learning methods become important, since new training data and categories will arrive over time. We will develop online learning methods that can incorporate new examples for existing category models, and learn new category models from few examples by leveraging similarity to related categories using multi-task learning methods. Here we will develop new distance-based classifiers and attribute and label embedding techniques, and explore the use of NLP techniques such as skipgram models to automatically determine between which classes transfer should occur. Moreover, NLP will be useful in the context of learning models for many categories to identify synonyms, and to determine cases of polysemy (e.g. jaguar car brand v.s. jaguar animal), and merge or refine categories accordingly. Ultimately this will result in methods that are able to learn an"encyclopedia" of visual models.
- **Visual search from unstructured textual queries.** We will build on recent approaches that learn recognition models on-the-fly (as the query is issued) from generic image search engines such as Google Images. While it is feasible to learn models in this manner in a matter of seconds, it is challenging to use the model to retrieve relevant content in real-time from large video archives of more than a few thousand hours. To achieve this requires feature compression techniques to store visual representations in memory, and cascaded search techniques to avoid exhaustive search. This approach, however, leaves untouched the core problem of how to associate visual material with the textual query in the first place. The second approach we will explore is based on image annotation models. In particular we will go beyond image-text retrieval methods by using recurrent neural networks such as Elman networks or long short-term memory (LSTM) networks to generate natural language sentences to describe images.

3.3. Large-scale learning and optimization

We have entered an era of massive data acquisition, leading to the revival of an old scientific utopia: it should be possible to better understand the world by automatically converting data into knowledge. It is also leading to a new economic paradigm, where data is a valuable asset and a source of activity. Therefore, developing scalable technology to make sense of massive data has become a strategic issue. Computer vision has already started to adapt to these changes.

In particular, very high dimensional models such as deep networks are becoming highly popular and successful for visual recognition. This change is closely related to the advent of big data. On the one hand, these models involve a huge number of parameters and are rich enough to represent well complex objects such as natural images or text corpora. On the other hand, they are prone to overfitting (fitting too closely to training data without being able to generalize to new unseen data) despite regularization; to work well on difficult tasks, they require a large amount of labelled data that has been available only recently. Other cues may explain their success: the deep learning community has made significant engineering efforts, making it possible to learn in a day on a GPU large models that would have required weeks of computations on a traditional CPU, and it has accumulated enough empirical experience to find good hyper-parameters for its networks.

To learn the huge number of parameters of deep hierarchical models requires scalable optimization techniques and large amounts of data to prevent overfitting. This immediately raises two major challenges: how to learn without large amounts of labeled data, or with weakly supervised annotations? How to efficiently learn such huge-dimensional models? To answer the above challenges, we will concentrate on the design and theoretical

justifications of deep architectures including our recently proposed deep kernel machines, with a focus on weakly supervised and unsupervised learning, and develop continuous and discrete optimization techniques that push the state of the art in terms of speed and scalability.

This research axis will be developed into three sub-tasks:

- **Deep kernel machines for structured data.** Deep kernel machines combine advantages of kernel methods and deep learning. Both approaches rely on high-dimensional models. Kernels implicitly operate in a space of possibly infinite dimension, whereas deep networks explicitly construct high-dimensional nonlinear data representations. Yet, these approaches are complementary: Kernels can be built with deep learning principles such as hierarchies and convolutions, and approximated by multilayer neural networks. Furthermore, kernels work with structured data and have well understood theoretical principles. Thus, a goal of the Thoth project-team is to design and optimize the training of such deep kernel machines.
- **Large-scale parallel optimization.** Deep kernel machines produce nonlinear representations of input data points. After encoding these data points, a learning task is often formulated as a *large-scale convex optimization problem*; for example, this is the case for linear support vector machines, logistic regression classifiers, or more generally many empirical risk minimization formulations. We intend to pursue recent efforts for making convex optimization techniques that are dedicated to machine learning more scalable. Most existing approaches address scalability issues either in model size (meaning that the function to minimize is defined on a domain of very high dimension), or in the amount of training data (typically, the objective is a large sum of elementary functions). There is thus a large room for improvements for techniques that jointly take these two criteria into account.
- **Large-scale graphical models.** To represent structured data, we will also investigate graphical models and their optimization. The challenge here is two-fold: designing an adequate cost function and minimizing it. While several cost functions are possible, their utility will be largely determined by the efficiency and the effectiveness of the optimization algorithms for solving them. It is a combinatorial optimization problem involving billions of variables and is NP-hard in general, requiring us to go beyond the classical approximate inference techniques. The main challenges in minimizing cost functions stem from the large number of variables to be inferred, the inherent structure of the graph induced by the interaction terms (e.g., pairwise terms), and the high-arity terms which constrain multiple entities in a graph.

3.4. Datasets and evaluation

Standard benchmarks with associated evaluation measures are becoming increasingly important in computer vision, as they enable an objective comparison of state-of-the-art approaches. Such datasets need to be relevant for real-world application scenarios; challenging for state-of-the-art algorithms; and large enough to produce statistically significant results.

A decade ago, small datasets were used to evaluate relatively simple tasks, such as for example interest point matching and detection. Since then, the size of the datasets and the complexity of the tasks gradually evolved. An example is the Pascal Visual Object Challenge with 20 classes and approximately 10,000 images, which evaluates object classification and detection. Another example is the ImageNet challenge, including thousands of classes and millions of images. In the context of video classification, the TrecVid Multimedia Event Detection challenges, organized by NIST, evaluate activity classification on a dataset of over 200,000 video clips, representing more than 8,000 hours of video, which amounts to 11 months of continuous video.

Almost all of the existing image and video datasets are annotated by hand; it is the case for all of the above cited examples. In some cases, they present limited and unrealistic viewing conditions. For example, many images of the ImageNet dataset depict upright objects with virtually no background clutter, and they may not capture particularly relevant visual concepts: most people would not know the majority of subcategories of snakes cataloged in ImageNet. This holds true for video datasets as well, where in addition a taxonomy of action and event categories is missing.

Our effort on data collection and evaluation will focus on two directions. First, we will design and assemble video datasets, in particular for action and activity recognition. This includes defining relevant taxonomies of actions and activities. Second, we will provide data and define evaluation protocols for weakly supervised learning methods. This does not mean of course that we will forsake human supervision altogether: some amount of ground-truth labeling is necessary for experimental validation and comparison to the state of the art. Particular attention will be payed to the design of efficient annotation tools.

Not only do we plan to collect datasets, but also to provide them to the community, together with accompanying evaluation protocols and software, to enable a comparison of competing approaches for action recognition and large-scale weakly supervised learning. Furthermore, we plan to set up evaluation servers together with leaderboards, to establish an unbiased state of the art on held out test data for which the ground-truth annotations are not distributed. This is crucial to avoid tuning the parameters for a specific dataset and to guarantee a fair evaluation.

- **Action recognition.** We will develop datasets for recognizing human actions and human-object interactions (including multiple persons) with a significant number of actions. Almost all of today's action recognition datasets evaluate classification of short video clips into a number of predefined categories, in many cases a number of different sports, which are relatively easy to identify by their characteristic motion and context. However, in many real-world applications the goal is to identify and localize actions in entire videos, such as movies or surveillance videos of several hours. The actions targeted here are "real-world" and will be defined by compositions of atomic actions into higher-level activities. One essential component is the definition of relevant taxonomies of actions and activities. We think that such a definition needs to rely on a decomposition of actions into poses, objects and scenes, as determining all possible actions without such a decomposition is not feasible. We plan to provide annotations for spatio-temporal localization of humans as well as relevant objects and scene parts for a large number of actions and videos.
- **Weakly supervised learning.** We will collect weakly labeled images and videos for training. The collection process will be semi-automatic. We will use image or video search engines such as Google Image Search, Flickr or YouTube to find visual data corresponding to the labels. Initial datasets will be obtained by manually correcting whole-image/video labels, i.e., the approach will evaluate how well the object model can be learned if the entire image or video is labeled, but the object model has to be extracted automatically. Subsequent datasets will features noisy and incorrect labels. Testing will be performed on PASCAL VOC'07 and ImageNet, but also on more realistic datasets similar to those used for training, which we develop and manually annotate for evaluation. Our dataset will include both images and videos, the categories represented will include objects, scenes as well as human activities, and the data will be presented in realistic conditions.
- **Joint learning from visual information and text.** Initially, we will use a selection from the large number of movies and TV series for which scripts are available on-line, see for example http://www.dailyscript.com and http://www.weeklyscript.com. These scripts can easily be aligned with the videos by establishing correspondences between script words and (timestamped) spoken ones obtained from the subtitles or audio track. The goal is to jointly learn from visual content and text. To measure the quality of such a joint learning, we will manually annotate some of the videos. Annotations will include the space-time locations of the actions as well as correct parsing of the sentence. While DVDs will, initially, receive most attention, we will also investigate the use of data obtained from web pages, for example images with captions, or images and videos surrounded by text. This data is by nature more noisy than scripts.

4. Application Domains

4.1. Visual applications

Any solution to automatically understanding images and videos on a semantic level will have an immediate impact on a wide range of applications. For example:

- Semantic-level image and video access is highly relevant for visual search on the Web, in professional archives and personal collections.
- Visual data organization is applicable to organizing family photo and video albums as well as to large-scale information retrieval.
- Visual object recognition has potential applications ranging from surveillance, service robotics for assistance in day-to-day activities as well as the medical domain.
- Action recognition is highly relevant to visual surveillance, assisted driving and video access.
- Real-time scene understanding is relevant for human interaction through devices such as HoloLens, Oculus Rift.

4.2. Pluri-disciplinary research

Machine learning is intrinsically pluri-disciplinary. By developing large-scale machine learning models and algorithms for processing data, the Thoth team became naturally involved in pluri-disciplinary collaborations that go beyond visual modelling. In particular,

- extensions of unsupervised learning techniques originally developed for modelling the statistics of natural images have been deployed in neuro-imaging for fMRI data with the collaboration of the Parietal team from Inria.
- similarly, deep convolutional data representations, also originally developed for visual data, have been successfully extended to the processing of biological sequences, with collaborators from bioinformatics.
- Thoth also collaborates with experts in natural language and text processing, for applications where visual modalities need to be combined with text data.

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

- Cordelia Schmid was elected to the German National Academy of Sciences, Leopoldina, in 2017.
- Cordelia Schmid was a Highly Cited Researcher in 2017 (Clarivate Analytics former Thomson Reuters).
- Julien Mairal received the IEEE PAMI young researcher award.
- Gregory Rogez and Cordelia Schmid received an Amazon Academic Research Award.
- Gregory Rogez received an CVPR 2017 outsanding reviewer award.

6. New Software and Platforms

6.1. ACT-detector

Action Tubelet Detector for Spatio-Temporal Action Localization
KEYWORDS: Spatio-temporal - Localisation - Video analysis - Motion detection - Object detection

FUNCTIONAL DESCRIPTION: Current state-of-the-art approaches for spatio-temporal action detection rely on detections at the frame level that are then linked or tracked across time. In this paper, we leverage the temporal continuity of videos instead of operating at the frame level. We propose the ACtion Tubelet detector (ACT-detector) that takes as input a sequence of frames and outputs tubelets, ie., sequences of bounding boxes with associated scores. The same way state-of-the-art object detectors rely on anchor boxes, our ACT-detector is based on anchor cuboids. We build upon the state-of-the-art SSD framework. Convolutional features are extracted for each frame, while scores and regressions are based on the temporal stacking of these features, thus exploiting information from a sequence. Our experimental results show that leveraging sequences of frames significantly improves detection performance over using individual frames. The gain of our tubelet detector can be explained by both more relevant scores and more precise localization. Our ACT-detector outperforms the state of the art methods for frame-mAP and video-mAP on the J-HMDB and UCF-101 datasets, in particular at high overlap thresholds.

- Participants: Philippe Weinzaepfel, Vittorio Ferrari, Cordelia Schmid and Vasiliki Kalogeiton
- Contact: Vasiliki Kalogeiton
- Publication: Action Tubelet Detector for Spatio-Temporal Action Localization
- URL: http://thoth.inrialpes.fr/src/ACTdetector/

6.2. Joint object-action learning

Joint learning of object and action detectors
KEYWORDS: Detection - Video sequences - Zero-shot
SCIENTIFIC DESCRIPTION: we propose to jointly detect object-action instances in uncontrolled videos, e.g. cat eating, dog running or car rolling. We build an end-to-end two stream network architecture for joint learning of objects and actions. We cast this joint learning problem by leveraging a multitask objective. We compare our proposed end-to-end multitask architecture with alternative ones: (i) treating every possible combination of actions and objects as a separate class (Cartesian) and (ii) considering a hierarchy of objects-actions: the first level corresponds to objects and the second one to the valid actions for each object (hierarchical). We show that our method performs as well as these two alternatives while (a) requiring fewer parameters and (b) enabling zero-shot learning of the actions performed by a specific object, i.e., when training for an object class alone without its actions, our multitask network is able to predict actions for that object class by leveraging actions performed by other objects. our multitask objective not only allows to effectively detect object-action pairs but also leads to performance improvements on each individual task (i.e., detection of either objects or actions). We compare to the state of the art for object-action detection on the Actor-Action (A2D) dataset and we outperform it.
FUNCTIONAL DESCRIPTION: While most existing approaches for detection in videos focus on objects or human actions separately, we aim at jointly detecting objects performing actions, such as cat eating or dog jumping. We introduce an end-to-end multitask objective that jointly learns object-action relationships. We compare it with different training objectives, validate its effectiveness for detecting objects-actions in videos, and show that both tasks of object and action detection benefit from this joint learning. Moreover, the proposed architecture can be used for zero-shot learning of actions: our multitask objective leverages the commonalities of an action performed by different objects, e.g. dog and cat jumping , enabling to detect actions of an object without training with these object-actions pairs. In experiments on the A2D dataset, we obtain state-of-the-art results on segmentation of object-action pairs. We finally apply our multitask architecture to detect visual relationships between objects in images of the VRD dataset.

- Participants: Vasiliki Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari and Cordelia Schmid
- Contact: Vasiliki Kalogeiton
- Publication: Joint learning of object and action detectors
- URL: https://github.com/vkalogeiton/joint-object-action-learning

6.3. BlitzNet

A Real-Time Deep Network for Scene Understanding
KEYWORD: Computer vision
FUNCTIONAL DESCRIPTION: Real-time scene understanding has become crucial in many applications such as autonomous driving. This deep architecture, called BlitzNet, jointly performs object detection and semantic segmentation in one forward pass, allowing real-time computations. Besides the computational gain of having a single network to perform several tasks, object detection and semantic segmentation benefit from each other in terms of accuracy.

- Participants: Mikita Dvornik, Konstantin Shmelkov, Julien Mairal and Cordelia Schmid
- Contact: Mikita Dvornik
- Publication: BlitzNet: A Real-Time Deep Network for Scene Understanding
- URL: https://github.com/dvornikita/blitznet

6.4. LCR-Net

Localization-Classification-Regression Network for Human Pose
KEYWORDS: Object detection - Recognition of human movement
FUNCTIONAL DESCRIPTION: We propose an end-to-end architecture for joint 2D and 3D human pose estimation in natural images. Key to our approach is the generation and scoring of a number of pose proposals per image, which allows us to predict 2D and 3D pose of multiple people simultaneously. Our architecture contains 3 main components: 1) the pose proposal generator that suggests potential poses at different locations in the image, 2) a classifier that scores the different pose proposals , and 3) a regressor that refines pose proposals both in 2D and 3D.

- Participants: Grégory Rogez, Philippe Weinzaepfel and Cordelia Schmid
- Contact: Grégory Rogez
- Publication: LCR-Net: Localization-Classification-Regression for Human Pose
- URL: https://thoth.inrialpes.fr/src/LCR-Net/

6.5. CKN-seq

Convolutional Kernel Networks for Biological Sequences
KEYWORD: Bioinformatics
SCIENTIFIC DESCRIPTION: The growing amount of biological sequences available makes it possible to learn genotype-phenotype relationships from data with increasingly high accuracy. By exploiting large sets of sequences with known phenotypes, machine learning methods can be used to build functions that predict the phenotype of new, unannotated sequences. In particular, deep neural networks have recently obtained good performances on such prediction tasks, but are notoriously difficult to analyze or interpret. Here, we introduce a hybrid approach between kernel methods and convolutional neural networks for sequences, which retains the ability of neural networks to learn good representations for a learning problem at hand, while defining a well characterized Hilbert space to describe prediction functions. Our method outperforms state-of-the-art convolutional neural networks on a transcription factor binding prediction task while being much faster to train and yielding more stable and interpretable results.
FUNCTIONAL DESCRIPTION: CKN-Seq is a software package for predicting transcription factor binding sites. It was shipped with the BiorXiv preprint

D. Chen, L. Jacob, and J. Mairal. Predicting Transcription Factor Binding Sites with Convolutional Kernel Networks. 2017.

The software is implemented in PyTorch.

- Participants: Laurent Jacob, Dexiong Chen and Julien Mairal
- Partners: CNRS - UGA
- Contact: Julien Mairal
- Publication: Predicting Transcription Factor Binding Sites with Convolutional Kernel Networks
- URL: https://gitlab.inria.fr/dchen/CKN-seq

6.6. CKN-TensorFlow

Convolutional Kernel Networks in TensorFlow
KEYWORD: Machine learning
SCIENTIFIC DESCRIPTION: This software package implements a new image representation based on a multilayer kernel machine. Unlike traditional kernel methods where data representation is decoupled from the prediction task, we learn how to shape the kernel with supervision. We proceed by first proposing improvements of the recently-introduced convolutional kernel networks (CKNs) in the context of unsupervised learning, then, we derive backpropagation rules to take advantage of labeled training data. The resulting model is a new type of convolutional neural network, where optimizing the filters at each layer is equivalent to learning a linear subspace in a reproducing kernel Hilbert space (RKHS).
FUNCTIONAL DESCRIPTION: This is the implementation in TensorFlow of the Convolutional Kernel Networks method for image classification, described in the paper J. Mairal. End-to-End Kernel Learning with Supervised Convolutional Kernel Networks. Adv. Neural Information Processing Systems (NIPS), 2016.

- Participants: Ghislain Durif and Julien Mairal
- Contact: Ghislain Durif
- Publication: End-to-End Kernel Learning with Supervised Convolutional Kernel Networks

6.7. Stochs

fast stochastic solvers for machine learning
KEYWORD: Machine learning
FUNCTIONAL DESCRIPTION: The stochs library provides efficient C++ implementations of stochastic optimization algorithms for common machine learning settings, including situations with finite datasets augmented with random perturbations (e.g. data augmentation or dropout). The library is mainly used from Python through a Cython extension. Currently, SGD, (S-)MISO and (N-)SAGA are supported, for dense and sparse data. See the following reference for details:

A. Bietti and J. Mairal. Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite-Sum Structure. arXiv 1610.00970, 2017.

- Participants: Alberto Bietti and Julien Mairal
- Contact: Alberto Bietti
- Publication: Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite-Sum Structure
- URL: https://github.com/albietz/stochs

6.8. MODL

Massive Online Dictionary Learning
KEYWORDS: Pattern discovery - Machine learning

FUNCTIONAL DESCRIPTION: Matrix factorization library, usable on very large datasets, with optional sparse and positive factors.

- Participants: Arthur Mensch, Gaël Varoquaux, Bertrand Thirion and Julien Mairal
- Contact: Arthur Mensch
- Publications: Subsampled online matrix factorization with convergence guarantees - Stochastic Subsampling for Factorizing Huge Matrices
- URL: http://github.com/arthurmensch/modl

6.9. Loter

Loter: A software package to infer local ancestry for a wide range of species
KEYWORDS: Local Ancestry Inference - Bioinformatics
SCIENTIFIC DESCRIPTION: Admixture between populations provides opportunity to study biological adaptation and phenotypic variation. Admixture studies can rely on local ancestry inference for admixed individuals, which consists of computing at each locus the number of copies that originate from ancestral source populations. Loter is a software package that does not require any biological parameter besides haplotype data in order to make local ancestry inference available for a wide range of species.
FUNCTIONAL DESCRIPTION: Loter is a Python package for haplotype phasing and local ancestry inference.
NEWS OF THE YEAR: The software package was shipped with the biorxiv preprint T. Dias-Alves, J. Mairal, and M. Blum. Loter: A Software Package to Infer Local Ancestry for a Wide Range of Species. preprint BiorXiv. 2017

- Participants: Thomas Dias-Alves, Michael Blum and Julien Mairal
- Partners: UGA - CNRS
- Contact: Julien Mairal
- Publication: Loter: A software package to infer local ancestry for a wide range of species
- URL: https://github.com/bcm-uga/Loter

6.10. SPAMS

SPArse Modeling Software
KEYWORDS: Signal processing - Machine learning
FUNCTIONAL DESCRIPTION: SPAMS is an open-source software package for sparse estimation
NEWS OF THE YEAR: The version 2.6.1 of the software package is now compatible with Python v3, R v3, comes with pre-compiled Matlab packages, and is now available on the conda and PyPi package managers.

- Participants: Ghislain Durif and Julien Mairal
- Contact: Julien Mairal
- URL: http://spams-devel.gforge.inria.fr/

6.11. MP-Net

KEYWORD: Motion analysis
FUNCTIONAL DESCRIPTION: This is a public implementation of the method described in the following paper: Learning Motion Patterns in Videos [CVPR 2017].

The problem of determining whether an object is in motion, irrespective of the camera motion, is far from being solved. We address this challenging task by learning motion patterns in videos. The core of our approach is a fully convolutional network, which is learnt entirely from synthetic video sequences, and their ground-truth optical flow and motion segmentation. This encoder-decoder style architecture first learns a coarse representation of the optical flow field features, and then refines it iteratively to produce motion labels at the original high-resolution. The output label of each pixel denotes whether it has undergone independent motion, i.e., irrespective of the camera motion. We demonstrate the benefits of this learning framework on the moving object segmentation task, where the goal is to segment all the objects in motion. To this end we integrate an objectness measure into the framework. Our approach outperforms the top method on the recently released DAVIS benchmark dataset, comprising real-world sequences, by 5.6

- Participants: Pavel Tokmakov, Karteek Alahari and Cordelia Schmid
- Contact: Pavel Tokmakov
- Publication: Learning Motion Patterns in Videos
- URL: http://thoth.inrialpes.fr/research/mpnet/

6.12. LVO

Learning Video Object Segmentation with Visual Memory
KEYWORD: Video analysis
FUNCTIONAL DESCRIPTION: This is a public implementation of the method described in the following paper: Learning Video Object Segmentation with Visual Memory [ICCV 2017].

This paper addresses the task of segmenting moving objects in unconstrained videos. We introduce a novel two-stream neural network with an explicit memory module to achieve this. The two streams of the network encode spatial and temporal features in a video sequence respectively, while the memory module captures the evolution of objects over time. The module to build a "visual memory" in video, i.e., a joint representation of all the video frames, is realized with a convolutional recurrent unit learned from a small number of training video sequences. Given a video frame as input, our approach assigns each pixel an object or background label based on the learned spatio-temporal features as well as the "visual memory" specific to the video, acquired automatically without any manually-annotated frames. The visual memory is implemented with convolutional gated recurrent units, which allows to propagate spatial information over time. We evaluate our method extensively on two benchmarks, DAVIS and Freiburg-Berkeley motion segmentation datasets, and show state-of-the-art results. For example, our approach outperforms the top method on the DAVIS dataset by nearly 6

- Participants: Karteek Alahari, Cordelia Schmid and Pavel Tokmakov
- Contact: Pavel Tokmakov
- Publication: Learning Video Object Segmentation with Visual Memory
- URL: http://lear.inrialpes.fr/research/lvo/

6.13. 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/

7. New Results

7.1. Visual recognition in images

7.1.1. Dynamic Filters in Graph Convolutional Networks

Participants: Nitika Verma, Edmond Boyer [MORPHEO, Inria Grenoble], Jakob Verbeek.

Convolutional neural networks (CNNs) have massively impacted visual recognition in 2D images, and are now ubiquitous in state-of-the-art approaches. While CNNs naturally extend to other domains, such as audio and video, where data is also organized in rectangular grids, they do not easily generalize to other types of data such as 3D shape meshes, social network graphs or molecular graphs. In our recent paper [39], we propose a novel graph-convolutional network architecture to handle such data. The architecture builds on a generic formulation that relaxes the 1-to-1 correspondence between filter weights and data elements around the center of the convolution, see Figure 1 for an illustration. The main novelty of our architecture is that the shape of the filter is a function of the features in the previous network layer, which is learned as an integral part of the neural network. Experimental evaluations on digit recognition, semi-supervised document classification, and 3D shape correspondence yield state-of-the-art results, significantly improving over previous work for shape correspondence.

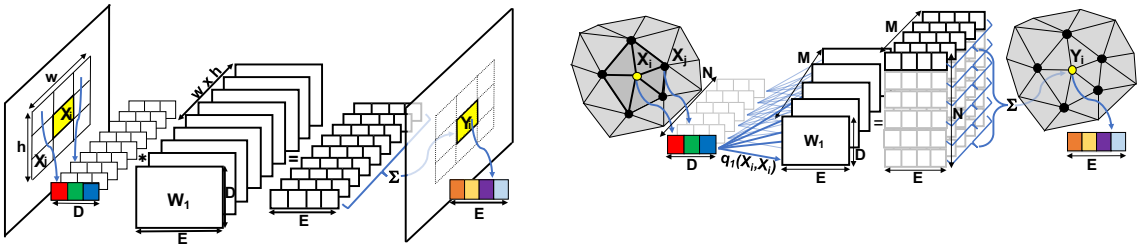

Figure 1. Left: Illustration of a standard CNN, representing the parameters as a set of $M = w \times h$ weight matrices, each of size $D \times E$. Each weight matrix is associated with a single relative position in the input patch. Right: Our graph convolutional network, where each node in the input patch is associated in a soft manner to each of the M weight matrices based on its features using the weight $q_m(\mathbf{x}_i, \mathbf{x}_j)$.

7.1.2. LCR-Net: Localization-Classification-Regression for Human Pose

Participants: Grégory Rogez, Philippe Weinzaepfel, Cordelia Schmid.

In this paper [24], we propose an end-to-end architecture for joint 2D and 3D human pose estimation in natural images. Key to our approach is the generation and scoring of a number of pose proposals per image, which allows us to predict 2D and 3D pose of multiple people simultaneously. See example in Figure 2. Hence, our approach does not require an approximate localization of the humans for initialization. Our architecture, named LCR-Net, contains 3 main components: 1) the pose proposal generator that suggests potential poses at different locations in the image; 2) a classifier that scores the different pose proposals ; and 3) a regressor that refines pose proposals both in 2D and 3D. All three stages share the convolutional feature layers and are trained jointly. The final pose estimation is obtained by integrating over neighboring pose hypotheses , which is shown to improve over a standard non maximum suppression algorithm. Our approach significantly outperforms the state of the art in 3D pose estimation on Human3.6M, a controlled environment. Moreover, it shows promising results on real images for both single and multi-person subsets of the MPII 2D pose benchmark.

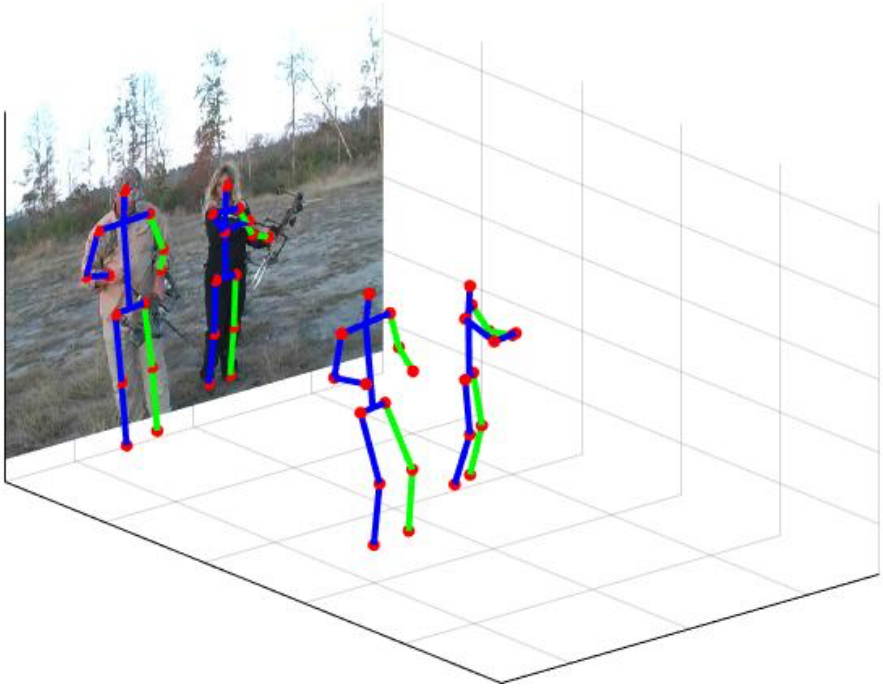

Figure 2. Examples of joint 2D-3D pose detections in a natural image. Even in case of occlusion or truncation, we estimate the joint locations by reasoning in term of full-body 2D-3D poses.

7.1.3. Incremental Learning of Object Detectors without Catastrophic Forgetting

Participants: Konstantin Shmelkov, Cordelia Schmid, Karteek Alahari.

In the paper [25] we introduce a framework for incremental learning of object detectors based on convolutional neural networks, i.e., adapting the original model trained on a set of classes to additionally detect objects of new classes, in the absence of the initial training data. They suffer from "catastrophic forgetting"—an abrupt degradation of performance on the original set of classes, when the training objective is adapted to the new classes. We present a method to address this issue, and learn object detectors incrementally, when neither the original training data nor annotations for the original classes in the new training set are available. The core of our proposed solution is a loss function to balance the interplay between predictions on the new classes and a new distillation loss which minimizes the discrepancy between responses for old classes from the original and the updated networks (see Figure 3). This incremental learning can be performed multiple times, for a new set of classes in each step, with a moderate drop in performance compared to the baseline network trained on the ensemble of data. We present object detection results on the PASCAL VOC 2007 and COCO datasets, along with a detailed empirical analysis of the approach.

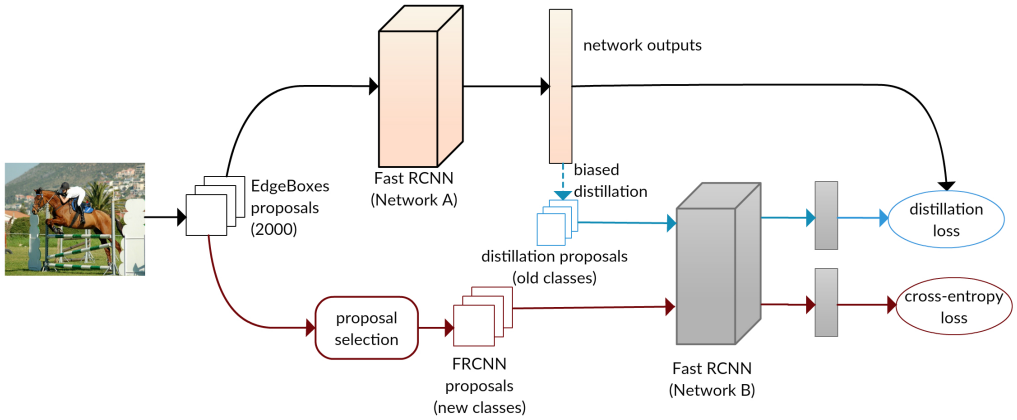

Figure 3. Overview of our framework for learning object detectors incrementally. It is composed of a frozen copy of the detector (Network A) and the detector (Network B) adapted for the new class(es).

7.1.4. BlitzNet: A Real-Time Deep Network for Scene Understanding

Participants: Mikita Dvornik, Konstantin Shmelkov, Julien Mairal, Cordelia Schmid.

Real-time scene understanding has become crucial in many applications such as autonomous driving. In this work [16], we propose a deep architecture, called BlitzNet, that jointly performs object detection and semantic segmentation in one forward pass, allowing real-time computations. Besides the computational gain of having a single network to perform several tasks, we show that object detection and semantic segmentation benefit from each other in terms of accuracy. Experimental results for VOC and COCO datasets show state-of-the-art performance for object detection and segmentation among real time systems.
To achieve these goals we designed a novel architecture (see fig.4 that naturally suits well for each of the tasks by allowing embedding of precise local details and reach global semantical information in a single feature-space. This solution allows to better localize and segment small objects. The usage of common architecture for both tasks allows more efficient feature sharing and and a simple training procedure that introduces benefits for semantic segmentation by adding extra data with only bounding box annotations. To reduce the computational overhead introduced by the upscale stream we slightly modify the NMS procedure to speed up post-processing in test time without no effect on detection accuracy.

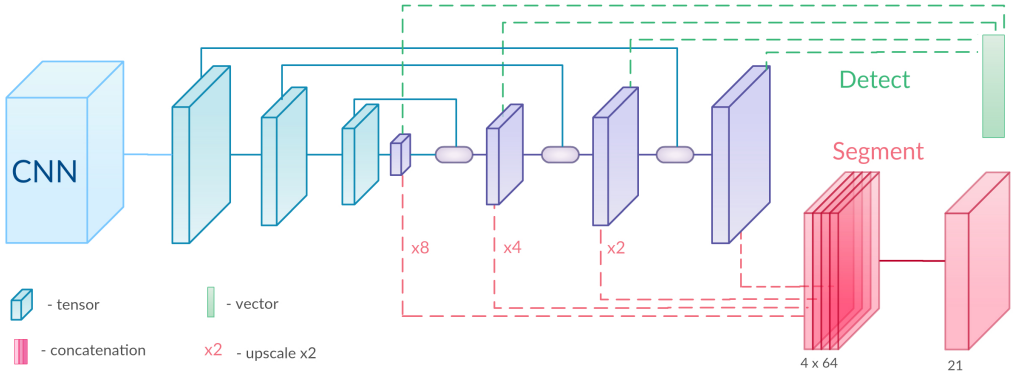

Figure 4. Atchirecture of the neutal network used in the project. The middle stream (blue and violet blocks) is shared between the task. The upper stream (in green) predicts categories of object proposals and their localization offsets to perform object detection. The bottom stream classifies each pixel to output a semantic segmentation mask.

7.1.5. SCNet: Learning semantic correspondence

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

In this work [17], 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 by 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.

7.1.6. Auxiliary Guided Autoregressive Variational Autoencoders

Participants: Thomas Lucas, Jakob Verbeek.

Generative modeling of high-dimensional data is a key problem in machine learning. Successful approaches include latent variable models and autoregressive models. The complementary strengths of these approaches, to model global and local image statistics respectively, suggest hybrid models combining the strengths of both. Our contribution is to train such hybrid models using an auxiliary loss function that controls which information is captured by the latent variables and what is left to the autoregressive decoder, as illustrated in Figure 6 . In contrast, prior work on such hybrid models needed to limit the capacity of the autoregressive decoder to prevent degenerate models that ignore the latent variables and only rely on autoregressive modeling. Our approach results in models with meaningful latent variable representations, and which rely on powerful autoregressive decoders to model image details. Our model generates qualitatively convincing samples, and yields state-of-the-art quantitative results.

7.1.7. Areas of Attention for Image Captioning

Participants: Marco Pedersoli, Thomas Lucas, Jakob Verbeek.

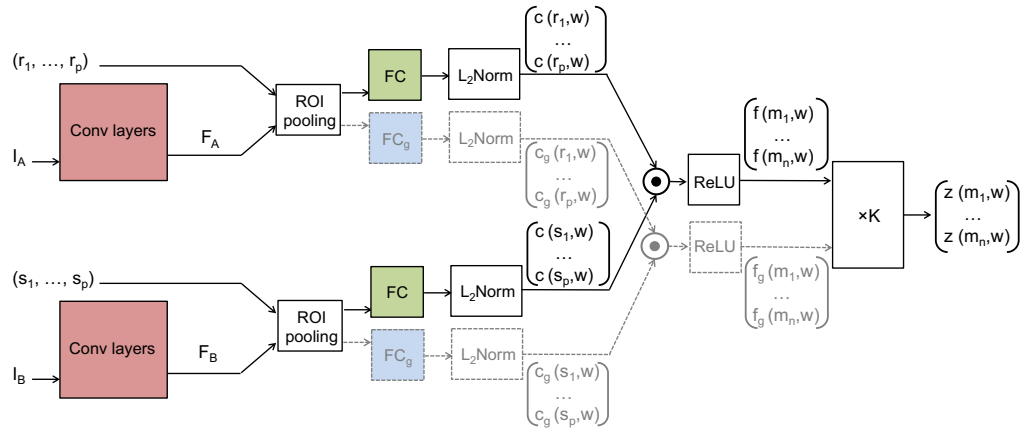

Figure 5. The SCNet architectures. Three variants are proposed: SCNet-AG, SCNet-A, and SCNet-AG+. The basic architecture, SCNet-AG, is drawn in solid lines. Colored boxes represent layers with learning parameters and the boxes with the same color share the same parameters. "$\times K$" denotes the voting 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.

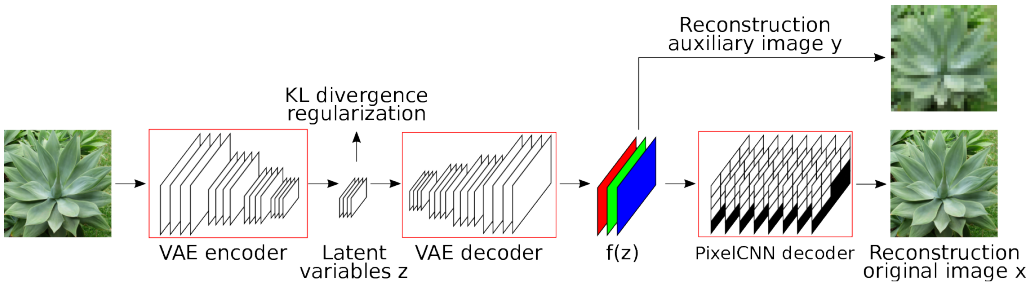

Figure 6. Schematic illustration of our auxiliary guided autoregressive variational autoencoder (AGAVE). An input image is encoded into a latent representation and decoded back into an image. This first reconstruction is guided by an auxiliary maximum likelihood loss and regularized with a Kullback-Liebler divergence. An autoregressive model is then conditionned on the auxiliary reconstruction and also trained with maximum likelihood.

We propose “Areas of Attention”, a novel attention-based model for automatic image captioning. Our approach models the dependencies between image regions, caption words, and the state of an RNN language model, using three pairwise interactions. In contrast to previous attention-based approaches that associate image regions only to the RNN state, our method allows a direct association between caption words and image regions. During training these associations are inferred from image-level captions, akin to weakly-supervised object detector training. These associations help to improve captioning by localizing the corresponding regions during testing. We also propose and compare different ways of generating attention areas: CNN activation grids, object proposals, and spatial transformers nets applied in a convolutional fashion, as illustrated in Figure 7. Spatial transformers give the best results. They allow for image specific attention areas, and can be trained jointly with the rest of the network. Our attention mechanism and spatial transformer attention areas together yield state-of-the-art results on the MSCOCO dataset.

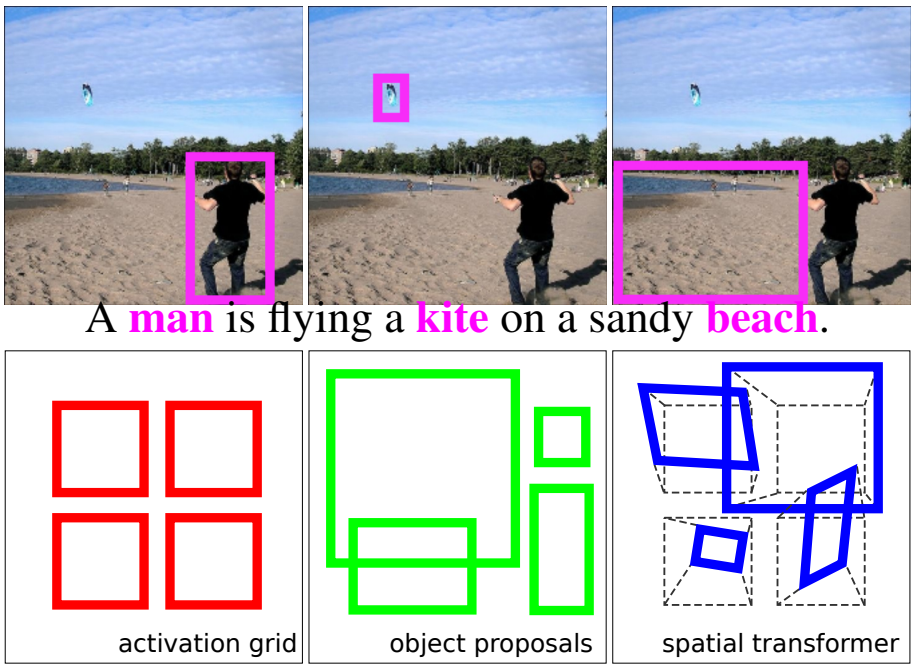

Figure 7. An attention mechanism jointly predicts the next caption word and the corresponding region at each time-step given the RNN state (top). Attention areas can be defined using CNN activation grids or object proposals (left and middle), as used in previous work. We also present a end-to-end trainable convolutional spatial transformer approach to compute image specific attention areas (bottom).

7.1.8. Enhancing Energy Minimization Framework for Scene Text Recognition with Top-Down Cues

Participants: Anand Mishra, Karteek Alahari, C. v. Jawahar.

Color and strokes are the salient features of text regions in an image. In this work, presented in [10], we use both these features as cues, and introduce a novel energy function to formulate the text binarization problem. The minimum of this energy function corresponds to the optimal binarization. We minimize the energy function with an iterative graph cut based algorithm. Our model is robust to variations in foreground and background as we learn Gaussian mixture models for color and strokes in each iteration of the graph cut. We show results on word images from the challenging ICDAR 2003/2011, born-digital image and street

view text datasets, as well as full scene images containing text from ICDAR 2013 datasets, such as the ones shown in Figure 8, and compare our performance with state-of-the-art methods. Our approach shows significant improvements in performance under a variety of performance measures commonly used to assess text binarization schemes. In addition, our method adapts to diverse document images, like text in videos, handwritten text images.

Figure 8. Sample images we consider in the work presented in [10]. Due to large variations in foreground and background colors, most of the popular binarization techniques in the literature tend to fail on such images.

7.1.9. Learning deep face representations using small data

Participants: Guosheng Hu, Xiaojiang Peng [Hengyang Normal University, China], Yongxin Yang [Queen Mary University of London, UK], Timothy Hospedales [University of Edinburgh, UK], Jakob Verbeek.

Deep convolutional neural networks have recently proven extremely effective for difficult face recognition problems in uncontrolled settings. To train such networks very large training sets are needed with millions of labeled images. For some applications, such as near-infrared (NIR) face recognition, such large training datasets are, however, not publicly available and very difficult to collect. In our recent paper [8] we propose a method to generate very large training datasets of synthetic images by compositing real face images in a given dataset. Our approach replaces facial parts (nose, mouth, eyes) from one face with those of another, see Figure 9 for several examples. We show that this method enables to learn models from as few as 10,000 training images, which perform on par with models trained from 500,000 images without using our data augmentation. Using our approach we also improve the state-of-the-art results on the CASIA NIR-VIS heterogeneous face recognition dataset.

7.1.10. Invariance and Stability of Deep Convolutional Representations

Participants: Alberto Bietti, Julien Mairal.

In [13] and [29], we study deep signal representations that are near-invariant to groups of transformations and stable to the action of diffeomorphisms without losing signal information. This is achieved by generalizing the multilayer kernel introduced in the context of convolutional kernel networks and by studying the geometry of the corresponding reproducing kernel Hilbert space. We show that the signal representation is stable, and that models from this functional space, such as a large class of convolutional neural networks, may enjoy the same stability.

7.1.11. Weakly-supervised learning of visual relations

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

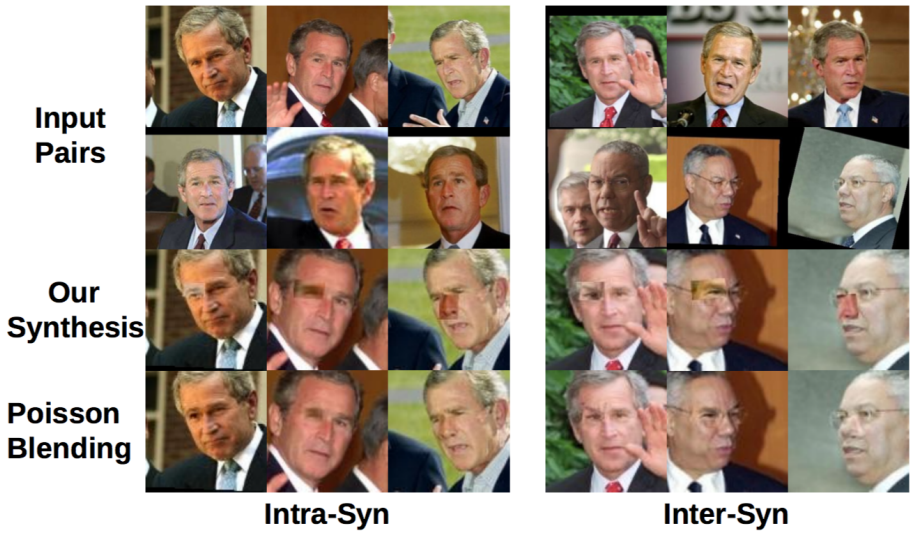

Figure 9. Illustration of generated synthetic training images, mixing either images of the same person (left) or from different people (right).

This work [23] 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. Example results are shown in Figure 10.

7.1.12. 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 [28], 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 allow for accurate human depth estimation and human part segmentation in real RGB images, see Figure 11. 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 at CVPR 2017 [28].

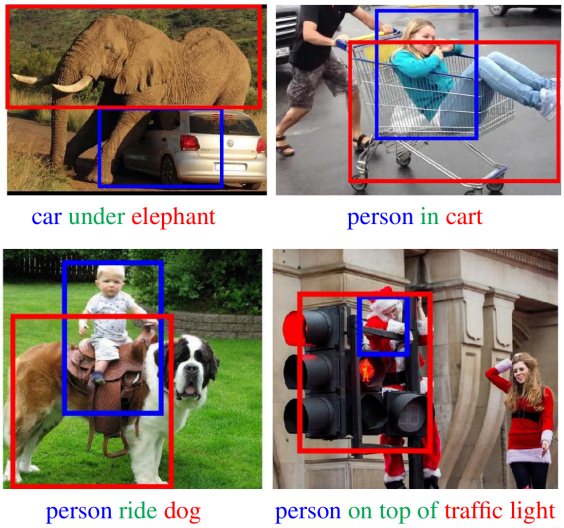

Figure 10. Examples of top retrieved pairs of boxes in UnRel dataset for unusual queries with our weakly supervised model

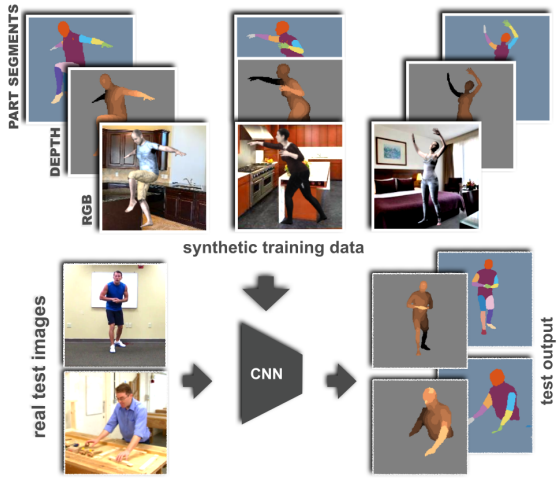

Figure 11. We generate photo-realistic synthetic images and their corresponding ground truth for learning pixel-wise classification problems: human parts segmentation and depth estimation. The convolutional neural network trained only on synthetic data generalizes on real images sufficiently for both tasks.

7.2. Visual recognition in videos

7.2.1. Detecting Parts for Action Localization

Participants: Nicolas Chesneau, Grégory Rogez, Karteek Alahari, Cordelia Schmid.

we propose a new framework for action localization that tracks people in videos and extracts full-body human tubes, i.e., spatio-temporal regions localizing actions, even in the case of occlusions or truncations. This is achieved by training a novel human part detector that scores visible parts while regressing full-body bounding boxes. The core of our method is a convolutional neural network which learns part proposals specific to certain body parts. These are then combined to detect people robustly in each frame. Our tracking algorithm connects the image detections temporally to extract full-body human tubes. We apply our new tube extraction method on the problem of human action localization, on the popular JHMDB dataset, and a very recent challenging dataset DALY (Daily Action Localization in YouTube), showing state-of-the-art results. An overview of the method is show in Figure 12. More details are provided in [15].

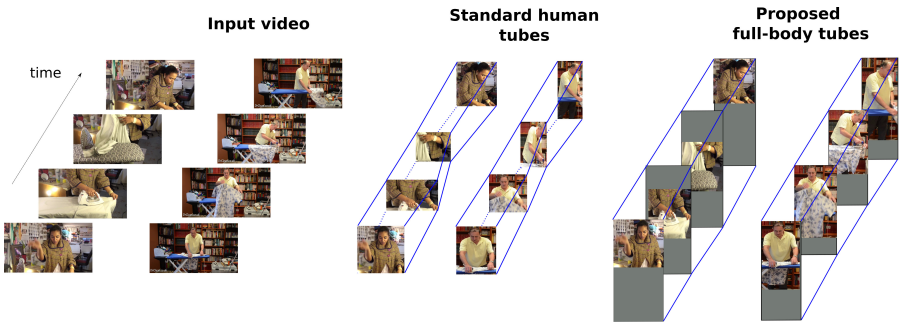

Figure 12. Two example videos from the DALY dataset to illustrate the difference between our human tube extraction and previous methods.

7.2.2. Learning from Web Videos for Event Classification

Participants: Nicolas Chesneau, Karteek Alahari, Cordelia Schmid.

Traditional approaches for classifying event videos rely on a manually curated training dataset. While this paradigm has achieved excellent results on benchmarks such as TrecVid multimedia event detection (MED) challenge datasets, it is restricted by the effort involved in careful annotation. Recent approaches have attempted to address the need for annotation by automatically extracting images from the web, or generating queries to retrieve videos. In the former case, they fail to exploit additional cues provided by video data, while in the latter, they still require some manual annotation to generate relevant queries. We take an alternate approach in this paper, leveraging the synergy between visual video data and the associated textual metadata, to learn event classifiers without manually annotating any videos. Specifically, we first collect a video dataset with queries constructed automatically from textual description of events, prune irrelevant videos with text and video data, and then learn the corresponding event classifiers. We evaluate this approach in the challenging setting where no manually annotated training set is available, i.e., EK0 in the TrecVid challenge, and show state-of-the-art results on MED 2011 and 2013 datasets. An overview of the method is show in Figure 13. More details are provided in [4].

7.2.3. Learning Motion Patterns in Videos

Participants: Pavel Tokmakov, Karteek Alahari, Cordelia Schmid.

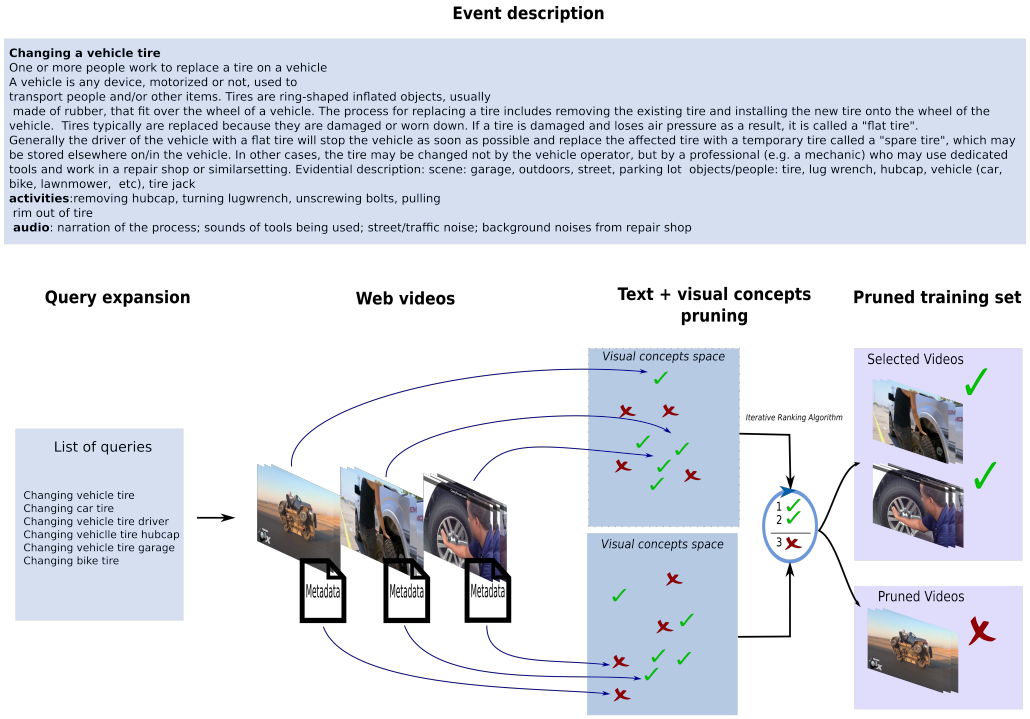

Figure 13. Overview: Given the description of an event ("Event description"), relevant queries are automatically generated ("Query generation") to collect an initial training set ("Web videos"). Text metadata and visual concepts extracted from these videos are used to select the relevant ones automatically ("Text + visual concepts pruning"), and build a training set for event classification ("Pruned training set").

The problem of determining whether an object is in motion, irrespective of the camera motion, is far from being solved. We address this challenging task by learning motion patterns in videos [26]. The core of our approach is a fully convolutional network (see Figure 14), which is learnt entirely from synthetic video sequences, and their ground-truth optical flow and motion segmentation. This encoder-decoder style architecture first learns a coarse representation of the optical flow field features, and then refines it iteratively to produce motion labels at the original high-resolution. The output label of each pixel denotes whether it has undergone independent motion, i.e., irrespective of the camera motion. We demonstrate the benefits of this learning framework on the moving object segmentation task, where the goal is to segment all the objects in motion. To this end we integrate an objectness measure into the framework. Our approach outperforms the top method on the recently released DAVIS benchmark dataset, comprising real-world sequences, by 5.6%. We also evaluate on the Berkeley motion segmentation database, achieving state-of-the-art results.

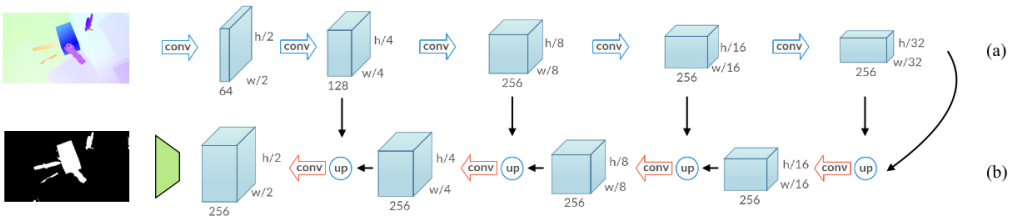

Figure 14. Our motion pattern network: MP-Net. The blue arrows in the encoder part (a) denote convolutional layers, together with ReLU and max-pooling layers. The red arrows in the decoder part (b) are convolutional layers with ReLU, 'up' denotes 2 x 2 upsampling of the output of the previous unit. The unit shown in green represents bilinear interpolation of the output of the last decoder unit

7.2.4. Learning Video Object Segmentation with Visual Memory

Participants: Pavel Tokmakov, Karteek Alahari, Cordelia Schmid.

This paper [27] addresses the task of segmenting moving objects in unconstrained videos. We introduce a novel two-stream neural network with an explicit memory module shown in Figure 15 to achieve this. The two streams of the network encode spatial and temporal features in a video sequence respectively, while the memory module captures the evolution of objects over time. The module to build a "visual memory" in video, i.e., a joint representation of all the video frames, is realized with a convolutional recurrent unit learned from a small number of training video sequences. Given a video frame as input, our approach assigns each pixel an object or background label based on the learned spatio-temporal features as well as the "visual memory" specific to the video, acquired automatically without any manually-annotated frames. The visual memory is implemented with convolutional gated recurrent units, which allows to propagate spatial information over time. We evaluate our method extensively on two benchmarks, DAVIS and Freiburg-Berkeley motion segmentation datasets, and show state-of-the-art results. For example, our approach outperforms the top method on the DAVIS dataset by nearly 6%. We also provide an extensive ablative analysis to investigate the influence of each component in the proposed framework.

7.2.5. Learning to Segment Moving Objects

Participants: Pavel Tokmakov, Cordelia Schmid, Karteek Alahari.

We study the problem of segmenting moving objects in unconstrained videos [37]. Given a video, the task is to segment all the objects that exhibit independent motion in at least one frame. We formulate this as a learning problem and design our framework with three cues: (i) independent object motion between a pair of frames, which complements object recognition, (ii) object appearance, which helps to correct errors in

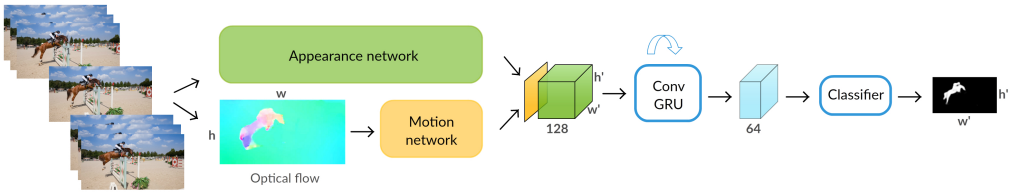

Figure 15. Overview of our segmentation approach. Each video frame is processed by the appearance (green) and the motion (yellow) networks to produce an intermediate two-stream representation. The ConvGRU module combines this with the learned visual memory to compute the final segmentation result.

motion estimation, and (iii) temporal consistency, which imposes additional constraints on the segmentation. The framework is a two-stream neural network with an explicit memory module, shown in Figure 15. The two streams encode appearance and motion cues in a video sequence respectively, while the memory module captures the evolution of objects over time, exploiting the temporal consistency. The motion stream is a convolutional neural network trained on synthetic videos to segment independently moving objects in the optical flow field. The module to build a "visual memory" in video, i.e., a joint representation of all the video frames, is realized with a convolutional recurrent unit learned from a small number of training video sequences. For every pixel in a frame of a test video, our approach assigns an object or background label based on the learned spatio-temporal features as well as the "visual memory" specific to the video. We evaluate our method extensively on three benchmarks, DAVIS, Freiburg-Berkeley motion segmentation dataset and SegTrack. In addition, we provide an extensive ablation study to investigate both the choice of the training data and the influence of each component in the proposed framework.

7.2.6. Joint learning of object and action detectors

Participants: Vasiliki Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, Cordelia Schmid.

While most existing approaches for detection in videos focus on objects or human actions separately, we aim at jointly detecting objects performing actions, such as cat eating or dog jumping [19]. We introduce an end-to-end multitask objective that jointly learns object-action relationships, see Figure 16. We compare it with different training objectives, validate its effectiveness for detecting objects-actions in videos, and show that both tasks of object and action detection benefit from this joint learning. Moreover, the proposed architecture can be used for zero-shot learning of actions: our multitask objective leverages the commonalities of an action performed by different objects, e.g. dog and cat jumping , enabling to detect actions of an object without training with these object-actions pairs. In experiments on the A2D dataset, we obtain state-of-the-art results on segmentation of object-action pairs. We finally apply our multitask architecture to detect visual relationships between objects in images of the VRD dataset.

7.2.7. Action Tubelet Detector for Spatio-Temporal Action Localization

Participants: Vasiliki Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, Cordelia Schmid.

Current state-of-the-art approaches for spatio-temporal action detection rely on detections at the frame level that are then linked or tracked across time. In this paper [18], we leverage the temporal continuity of videos instead of operating at the frame level. We propose the ACtion Tubelet detector (ACT-detector) that takes as input a sequence of frames and outputs tubelets, ie., sequences of bounding boxes with associated scores, see Figure 17. The same way state-of-the-art object detectors rely on anchor boxes, our ACT-detector is based on anchor cuboids. We build upon the state-of-the-art SSD framework. Convolutional features are extracted for each frame, while scores and regressions are based on the temporal stacking of these features, thus exploiting information from a sequence. Our experimental results show that leveraging sequences of frames

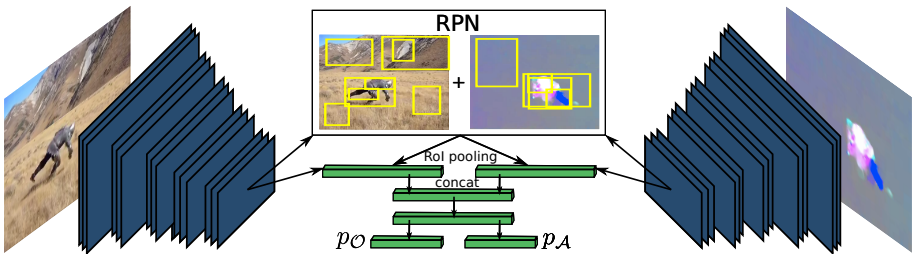

Figure 16. Overview of our end-to-end multitask network architecture for joint object-action detection in videos. Blue color represents convolutional layers while green represents fully connected layers. The end-to-end training is done by concatenating the fully connected layers from both streams. Here, pO and pA are the outputs of the two branches that predict the object and action labels, resulting in the final joint object-action loss.

significantly improves detection performance over using individual frames. The gain of our tubelet detector can be explained by both more relevant scores and more precise localization. Our ACT-detector outperforms the state of the art methods for frame-mAP and video-mAP on the J-HMDB and UCF-101 datasets, in particular at high overlap thresholds.

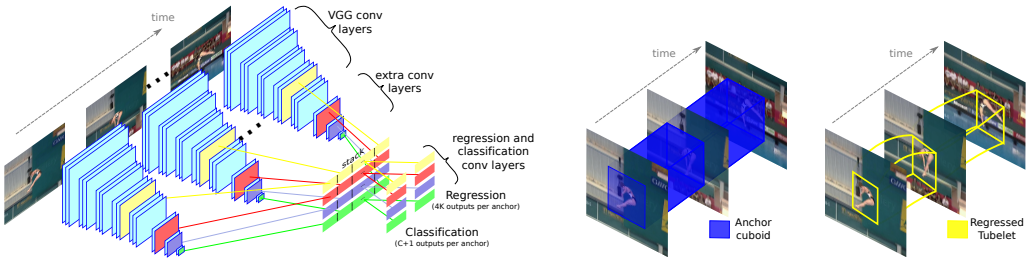

Figure 17. Overview of our ACT-detector. Given a sequence of frames, we extract convolutional features with weights shared between frames. We stack the features from subsequent frames to predict scores and regress coordinates for the anchor cuboids (middle figure, blue color). Depending on the size of the anchors, the features come from different convolutional layers (left figure, color coded: yellow, red, purple, green). As output, we obtain tubelets (right figure, yellow color).

7.3. Large-scale statistical learning

7.3.1. Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite-Sum Structure

Participants: Alberto Bietti, Julien Mairal.

Stochastic optimization algorithms with variance reduction have proven successful for minimizing large finite sums of functions. Unfortunately, these techniques are unable to deal with stochastic perturbations of input data, induced for example by data augmentation. In such cases, the objective is no longer a finite sum, and the main candidate for optimization is the stochastic gradient descent method (SGD). In [14], we introduce a variance reduction approach for these settings when the objective is composite and strongly convex. The

convergence rate outperforms SGD with a typically much smaller constant factor, which depends on the variance of gradient estimates only due to perturbations on a single example.

7.3.2. Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice

Participants: Hongzhou Lin, Julien Mairal, Zaid Harchaoui.

In this paper [35], we introduce a generic scheme for accelerating gradient-based optimization methods in the sense of Nesterov. The approach, called Catalyst, builds upon the inexact accelerated proximal point algorithm for minimizing a convex objective function, and consists of approximately solving a sequence of well-chosen auxiliary problems, leading to faster convergence. One of the key to achieve acceleration in theory and in practice is to solve these sub-problems with appropriate accuracy by using the right stopping criterion and the right warm-start strategy. In this paper, we give practical guidelines to use Catalyst and present a comprehensive theoretical analysis of its global complexity. We show that Catalyst applies to a large class of algorithms, including gradient descent, block coordinate descent, incremental algorithms such as SAG, SAGA, SDCA, SVRG, Finito/MISO, and their proximal variants. For all of these methods, we provide acceleration and explicit support for non-strongly convex objectives. We conclude with extensive experiments showing that acceleration is useful in practice, especially for ill-conditioned problems.

7.3.3. A Generic Quasi-Newton Algorithm for Faster Gradient-Based Optimization

Participants: Hongzhou Lin, Julien Mairal, Zaid Harchaoui.

In this paper [34], we propose a generic approach to accelerate gradient-based optimization algorithms with quasi-Newton principles. The proposed scheme, called QuickeNing, can be applied to incremental first-order methods such as stochastic variance-reduced gradient (SVRG) or incremental surrogate optimization (MISO). It is also compatible with composite objectives, meaning that it has the ability to provide exactly sparse solutions when the objective involves a sparsity-inducing regularization. QuickeNing relies on limited-memory BFGS rules, making it appropriate for solving high-dimensional optimization problems. Besides, it enjoys a worst-case linear convergence rate for strongly convex problems. We present experimental results where QuickeNing gives significant improvements over competing methods for solving large-scale high-dimensional machine learning problems, see Figure 18 for example.

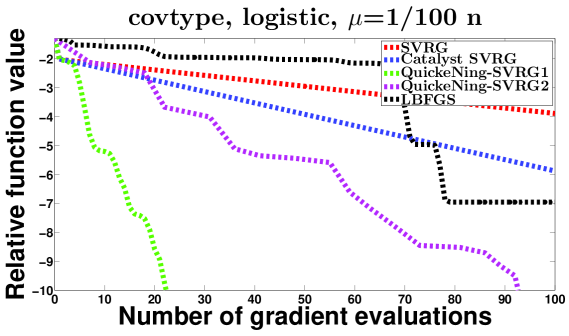

Figure 18. An illustration of the minimization of logistic regression. Significant improvement is observed by applying QuickeNing.

7.3.4. Catalyst Acceleration for Gradient-Based Non-Convex Optimization

Participants: Courtney Paquette, Hongzhou Lin, Dmitriy Drusvyatskiy, Julien Mairal, Zaid Harchaoui.

In this paper [36], we introduce a generic scheme to solve nonconvex optimization problems using gradient-based algorithms originally designed for minimizing convex functions. When the objective is convex, the proposed approach enjoys the same properties as the Catalyst approach of Lin et al, 2015. When the objective is nonconvex, it achieves the best known convergence rate to stationary points for first-order methods. Specifically, the proposed algorithm does not require knowledge about the convexity of the objective; yet, it obtains an overall worst-case efficiency of $O(\epsilon^{-2})$ and, if the function is convex, the complexity reduces to the near-optimal rate $O(\epsilon^{-2/3})$. We conclude the paper by showing promising experimental results obtained by applying the proposed approach to SVRG and SAGA for sparse matrix factorization and for learning neural networks (see Figure 19).

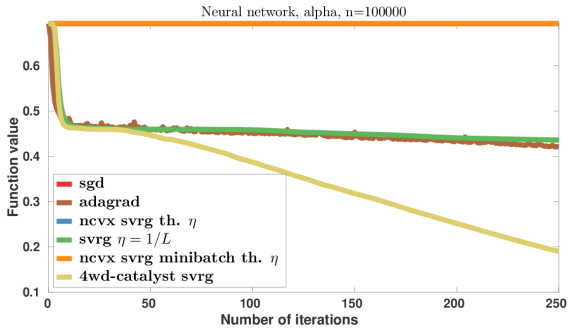

Figure 19. Comparison of different algorithms for the minimization of a two-layer neural network. Applying our method provides a clear acceleration in terms of function value.

7.4. Machine learning and pluri-disciplinary research

7.4.1. Predicting Transcription Factor Binding Sites with Convolutional Kernel Networks

Participants: Dexiong Chen, Laurent Jacob, Julien Mairal.

The growing amount of biological sequences available makes it possible to learn genotype-phenotype relationships from data with increasingly high accuracy. By exploiting large sets of sequences with known phenotypes, machine learning methods can be used to build functions that predict the phenotype of new, unannotated sequences. In particular, deep neural networks have recently obtained good performances on such prediction tasks, but are notoriously difficult to analyze or interpret. In this work, we introduce a hybrid approach between kernel methods and convolutional neural networks for sequences, which retains the ability of neural networks to learn good representations for a learning problem at hand, while defining a well characterized Hilbert space to describe prediction functions. Our method (see Figure 20), dubbed CKN-seq, outperforms state-of-the-art convolutional neural networks on a transcription factor binding prediction task while being much faster to train and yielding more stable and interpretable results.

Source code is freely available at https://gitlab.inria.fr/dchen/CKN-seq.

7.4.2. Loter: Inferring local ancestry for a wide range of species

Participants: Thomas Dias-Alves, Julien Mairal, Michael Blum [CNRS].

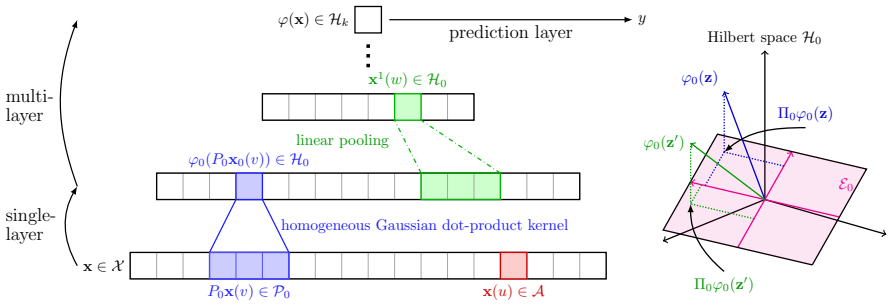

Figure 20. Construction scheme of CKN-seq

Admixture between populations provides opportunity to study biological adaptation and phenotypic variation. Admixture studies can rely on local ancestry inference for admixed individuals, which consists of computing at each locus the number of copies that originate from ancestral source populations. Existing software packages for local ancestry inference are tuned to provide accurate results on human data and recent admixture events. Here, we introduce Loter, an open-source software package that does not require any biological parameter besides haplotype data in order to make local ancestry inference available for a wide range of species. Using simulations, we compare the performance of Loter to HAPMIX, LAMP-LD, and RFMix. HAPMIX is the only software severely impacted by imperfect haplotype reconstruction. Loter is the less impacted software by increasing admixture time when considering simulated and admixed human genotypes. LAMP-LD and RFMIX are the most accurate method when admixture took place 20 generations ago or less; Loter accuracy is comparable or better than RFMix accuracy when admixture took place of 50 or more generations; and its accuracy is the largest when admixture is more ancient than 150 generations. For simulations of admixed Populus genotypes, Loter and LAMP-LD are robust to increasing admixture times by contrast to RFMix. When comparing length of reconstructed and true ancestry tracts, Loter and LAMP-LD provide results whose accuracy is again more robust than RFMix to increasing admixture times. We apply Loter to admixed Populus individuals and lengths of ancestry tracts indicate that admixture took place around 100 generations ago.

The Loter software package and its source code are available at https://github.com/bcm-uga/Loter.

7.4.3. High Dimensional Classification with combined Adaptive Sparse PLS and Logistic Regression

Participant: Ghislain Durif.

The high dimensionality of genomic data calls for the development of specific classification methodologies, especially to prevent over-optimistic predictions. This challenge can be tackled by compression and variable selection, which combined constitute a powerful framework for classification, as well as data visualization and interpretation. However, current proposed combinations lead to unstable and non convergent methods due to inappropriate computational frameworks. We hereby propose a computationally stable and convergent approach for classification in high dimensional based on sparse Partial Least Squares (sparse PLS). In this work [6], we start by proposing a new solution for the sparse PLS problem that is based on proximal operators for the case of univariate responses. Then we develop an adaptive version of the sparse PLS for classification, called logit-SPLS, which combines iterative optimization of logistic regression and sparse PLS to ensure computational convergence and stability. Our results are confirmed on synthetic and experimental data. In particular we show how crucial convergence and stability can be when cross-validation is involved for calibration purposes. Using gene expression data we explore the prediction of breast cancer relapse (c.f. figure 21 for an example of data visualization). We also propose a multi-categorical version of our method, used

to predict cell-types based on single-cell expression data. Our approach is implemented in the plsgenomics R-package.

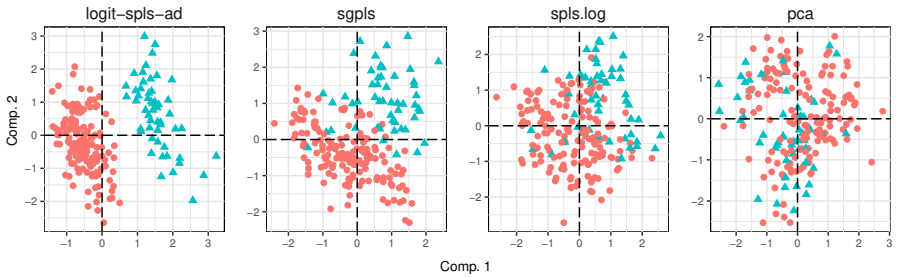

Figure 21. Visualization of gene expression profiles sampled from breast cancer patients in a two-dimensional subspace, using our supervised approach logit-SPLS, compared to other SPLS methods for supervised classification and PCA (unsupervised). Data are separated in two groups of individuals, presenting a relapse or not.

7.4.4. Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis

Participant: Ghislain Durif.

The development of high-throughput biology technologies now allows the investigation of the genome-wide diversity of transcription in single cells. This diversity has shown two faces: the expression dynamics (gene to gene variability) can be quantified more accurately, thanks to the measurement of lowly-expressed genes. Second, the cell-to-cell variability is high, with a low proportion of cells expressing the same gene at the same time/level. Those emerging patterns appear to be very challenging from the statistical point of view, especially to represent and to provide a summarized view of single-cell expression data. PCA is one of the most powerful framework to provide a suitable representation of high dimensional datasets, by searching for latent directions catching the most variability in the data. Unfortunately, classical PCA is based on Euclidean distances and projections that work poorly in presence of over-dispersed counts that show drop-out events (zero-inflation) like single-cell expression data. In this work [32], we propose a probabilistic Count Matrix Factorization (pCMF) approach for single-cell expression data analysis, that relies on a sparse Gamma-Poisson factor model. This hierarchical model is inferred using a variational EM algorithm. We show how this probabilistic framework induces a geometry that is suitable for single-cell data visualization, and produces a compression of the data that is very powerful for clustering purposes. Our method is competed to other standard representation methods like t-SNE, and we illustrate its performance for the representation of zero-inflated over-dispersed count data (c.f. figure 22). We also illustrate our work with results on a publicly available data set, being single-cell expression profile of neural stem cells. Our work is implemented in the pCMF R-package.

8. Bilateral Contracts and Grants with Industry

8.1. MSR-Inria joint lab: scientific image and video mining

Participants: Cordelia Schmid, Karteek Alahari.

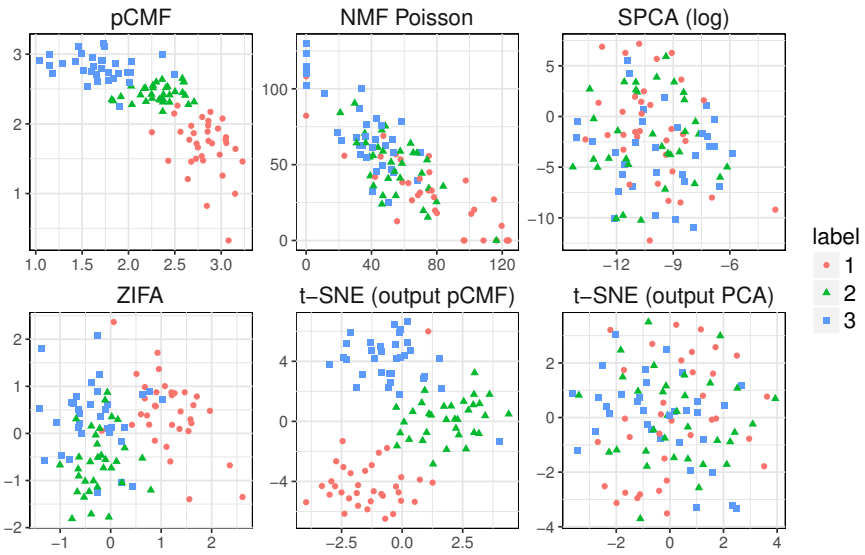

Figure 22. Visualization of synthetic zero-inflated over-dispersed count data in a two-dimensional subspace, using our approach pCMF, compared to PCA, Non-negative Matrix Factorization (NMF), Zero-Inflated Factor Analysis (ZIFA) and t-SNE. Data are generated with $n = 100$ indivuduals and $p = 1000$ recorded variables and 3 groups of individuals. t-SNE is applied with a preliminary dimension reduction step based on pCMF or PCA (default behavior).

This collaborative project, which started in September 2008, brings together the WILLOW and Thoth project-teams with researchers at Microsoft Research Cambridge and elsewhere. It 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 focuses on fundamental computer science research in computer vision and machine learning, and its application to archeology, cultural heritage preservation, environmental science, and sociology.

8.2. MSR-Inria joint lab: structured large-scale machine learning

Participants: Julien Mairal, Alberto Bietti, Hongzhou Lin.

Machine learning is now ubiquitous in industry, science, engineering, and personal life. While early successes were obtained by applying off-the-shelf techniques, there are two main challeges faced by machine learning in the " big data " era : structure and scale. The project proposes to explore three axes, from theoretical, algorithmic and practical perspectives: (1) large-scale convex optimization, (2) large-scale combinatorial optimization and (3) sequential decision making for structured data. The project involves two Inria sites and four MSR sites and started at the end of 2013.

8.3. Amazon

Participants: Grégory Rogez, Cordelia Schmid.

We received an Amazon Faculty Research Award end of 2016. The objective is 3D human action recognition from monocular RGB videos. The idea is to extend our recent work on human 3D pose estimation published at NIPS 2016 to videos and to develop an approach for action recognition based on temporal pose based on appropriate 3D features.

8.4. Intel

Participants: Cordelia Schmid, Karteek Alahari.

The Intel Network on Intelligent Systems in Europe brings together leading researchers in robotics, computer vision, motor control, and machine learning. We are part of this network and have participated in the annual retreat in 2017. Funding will be provided on an annual basis, every year, as long as we are part of the network.

8.5. Facebook

Participants: Cordelia Schmid, Jakob Verbeek, Karteek Alahari, Julien Mairal.

The collaboration started in 2016. The topics include image retrieval with CNN based descriptors, weakly supervised object detection and semantic segmentation, and learning structured models for action recognition in videos. In 2016, Pauline Luc started her PhD funded by a CIFRE grant, jointly supervised by Jakob Verbeek (Inria) and Camille Couprie (Facebook). THOTH has been selected in 2016 as a recipient for the Facebook GPU Partnership program. In this context Facebook has donated two state-of-the-art servers with 8 GPUs. In 2017, Alexandre Sablayrolles started his CIFRE grant, jointly supervised by Cordelia Schmid and Herve Jegou and Matthijs Douze at Facebook.

8.6. Xerox Research Center Europe

Participants: Cordelia Schmid, Vasileios Choutas, Philippe Weinzepfel [Naver].

The collaboration with Xerox has been on-going since October 2009 with two co-supervised CIFRE scholarships (2009–2012; 2011-2014). Starting June 2014 we signed a third collaborative agreement for a duration of three years. The goal is to develop approaches for deep learning based image description and pose estimation in videos. Jakob Verbeek (Inria) and Diane Larlus (XRCE) jointly supervise a PhD-level intern for a period of 6 months in 2016-2017. XRCE then became Naver in 2017 and the collaboration is still on-going, see next paragraph.

8.7. Naver

Participants: Karteek Alahari, Vladyslav Sydorov, Cordelia Schmid, Julien Mairal, Jakob Verbeek.

A one-year research contract on action recognition in videos started in Sept. 2017. The approach developed by V. Choutas implements pose-based motion features, which are shown to be complementary to state-of-the-art I3D features.

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. DeCore (Deep Convolutional and Recurrent networks for image, speech, and text)

Participants: Jakob Verbeek, Maha Elbayad.

DeCore is a project-team funded by the Persyval Lab for 3.5 years (september 2016 - February 2020), coordinated by Jakob Verbeek. It unites experts from Grenoble's applied-math and computer science labs LJK, GIPSA-LAB and LIG in the areas of computer vision, machine learning, speech, natural language processing, and information retrieval. The purpose of DeCore is to stimulate collaborative interdisciplinary research on deep learning in the Grenoble area, which is likely to underpin future advances in machine perception (vision, speech, text) over the next decade. It provides funding for two full PhD students. Maha Elbayad is one of them, supervised by Jakob Verbeek and Laurant Besacier (UGA).

9.2. National Initiatives

9.2.1. ANR Project Macaron

Participants: Julien Mairal, Zaid Harchaoui [University of Washington], Laurent Jacob [CNRS, LBBE Laboratory], Michael Blum [CNRS, TIMC Laboratory], Joseph Salmon [Telecom ParisTech], Mikita Dvornik, Thomas Dias-Alves, Daan Wynen.

The project MACARON is an endeavor to develop new mathematical and algorithmic tools for making machine learning more scalable. Our ultimate goal is to use data for solving scientific problems and automatically converting data into scientific knowledge by using machine learning techniques. Therefore, our project has two different axes, a methodological one, and an applied one driven by explicit problems. The methodological axis addresses the limitations of current machine learning for simultaneously dealing with large-scale data and huge models. The second axis addresses open scientific problems in bioinformatics, computer vision, image processing, and neuroscience, where a massive amount of data is currently produced, and where huge-dimensional models yield similar computational problems.

This is a 3 years and half project, funded by ANR under the program "Jeunes chercheurs, jeunes chercheuses", which started in October 2014. The principal investigator is Julien Mairal.

9.2.2. ANR Project DeepInFrance

Participant: Jakob Verbeek.

DeepInFrance (Machine learning with deep neural networks) project also aims at bringing together complementary machine learning, computer vision and machine listening research groups working on deep learning with GPUs in order to provide the community with the knowledge, the visibility and the tools that brings France among the key players in deep learning. The long-term vision of Deep in France is to open new frontiers and foster research towards algorithms capable of discovering sense in data in an automatic manner, a stepping stone before the more ambitious far-end goal of machine reasoning. The project partners are: INSA Rouen, Univ. Caen, Inria, UPMC, Aix-Marseille Univ., Univ. Nice Sophia Antipolis.

9.3. European Initiatives

9.3.1. FP7 & H2020 Projects

9.3.1.1. ERC Advanced grant Allegro

Participants: Cordelia Schmid, Pavel Tokmakov, Nicolas Chesneau, Vasiliki Kalogeiton, Konstantin Shmelkov, Daan Wynen, Xiaojiang Peng.

The ERC advanced grant ALLEGRO started in April 2013 for a duration of five years extended in 2017 for one year. The aim of ALLEGRO is to automatically learn from large quantities of data with weak labels. A massive and ever growing amount of digital image and video content is available today. It often comes with additional information, such as text, audio or other meta-data, that forms a rather sparse and noisy, yet rich and diverse source of annotation, ideally suited to emerging weakly supervised and active machine learning technology. The ALLEGRO project will take visual recognition to the next level by using this largely untapped source of data to automatically learn visual models. We will develop approaches capable of autonomously exploring evolving data collections, selecting the relevant information, and determining the visual models most appropriate for different object, scene, and activity categories. An emphasis will be put on learning visual models from video, a particularly rich source of information, and on the representation of human activities, one of today's most challenging problems in computer vision.

9.3.1.2. ERC Starting grant Solaris

Participants: Julien Mairal, Ghislain Durif, Andrei Kulunchakov, Dexiong Chen, Alberto Bietti, Hongzhou Lin.

The project SOLARIS started in March 2017 for a duration of five years. The goal of the project is to set up methodological and theoretical foundations of deep learning models, in the context of large-scale data processing. The main applications of the tools developed in this project are for processing visual data, such as videos, but also structured data produced in experimental sciences, such as biological sequences.

The main paradigm used in the project is that of kernel methods and consist of building functional spaces where deep learning models live. By doing so, we want to derive theoretical properties of deep learning models that may explain their success, and also obtain new tools with better stability properties. Another work package of the project is focused on large-scale optimization, which is a key to obtain fast learning algorithms.

9.4. International Initiatives

9.4.1. Inria Associate Teams Not Involved in an Inria International Labs

9.4.1.1. GAYA

Title: Semantic and Geometric Models for Video Interpretation

International Partner (Institution - Laboratory - Researcher):

Carnegie Mellon University (United States) - Robotics Institute - Deva Ramanan

Start year: 2016

See also: https://team.inria.fr/gaya/

The primary goal of the associate team GAYA is to interpret videos, in terms of recognizing actions, understanding the human-human and human-object interactions. Despite several years of research, it is yet unclear what is an efficient and robust video representation to attack this challenge. In order to address this, GAYA will focus on building semantic models, wherein we learn the video feature representation with limited supervision, and also geometric models, where we study the geometric properties of object shapes to better recognize them. The team consists of researchers from two Inria project-teams (Thoth and WILLOW) and a US university (Carnegie Mellon University [CMU]). It will allow the three teams to effectively combine their respective strengths in areas such as inference and machine learning approaches for vision tasks, feature representation, large-scale learning, geometric reasoning. The main expected outcomes of this collaboration are: effective

learnt representations of video content, new machine learning algorithms for handling minimally annotated data, large-scale public datasets for benchmarking, theoretical analysis of objects shapes and contours. In 2017, Gunnar Sigurdsson (PhD student of Abhinav Gupta [CMU]) visited the Thoth team to develop a new dataset of first- and third-person videos and an approach for learning a joint representation of these two modalities.

9.4.2. Inria International Partners

9.4.2.1. Informal International Partners

- **University of Edinburgh:** C. Schmid collaborates with V. Ferrari, full professor at university of Edinburgh. Vicky Kalogeiton started a co-supervised PhD in 2013 and graduated in 2017; she has been bi-localized between Uni. Edinburgh and Inria. Her subject is automatic learning of object representations in videos. The collaboration resulted in two joint publications in 2017 [19], [18].
- **MPI Tübingen:** C. Schmid collaborates with M. Black, a research director at MPI, starting in 2013. End of 2015 she was award a Humbolt research award funding a long-term research project with colleagues at MPI. She spent one month at MPI in May 2017. In 2017 the project resulted in the development of a large-scale synthetic human action dataset [12].
- **University of Washington:** Julien Mairal collaborates with Zaid Harchaoui, former member of the Lear team, on the topic of large-scale optimization. They co-advised one student, Hongzhou Lin, who defended his PhD in 2017.

9.4.3. Participation in Other International Programs

- **Indo-French project EVEREST** with IIIT Hyderabad, India, funded by CEFIPRA (Centre Franco-Indien pour la Promotion de la Recherche Avancee). The aim of this project between Cordelia Schmid, Karteek Alahari and C. V. Jawahar (IIIT Hyderabad) is to enable the use of rich, complex models that are required to address the challenges of high-level computer vision. The work plan for the project will follow three directions. First, we will develop a learning framework that can handle weak annotations. Second, we will build formulations to solve the non-convex optimization problem resulting from the learning framework. Third, we will develop efficient and accurate energy minimization algorithms, in order to make the optimization computationally feasible.

9.5. International Research Visitors

9.5.1. Visits to International Teams

9.5.1.1. Research Stays Abroad

- A. Bietti visited Microsoft Research at New York from September to December 2017, as part of the MSR-Inria joint centre collaboration.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. Member of the Organizing Committees

- C. Schmid. General Chair European Conference on Computer Vision (ECCV) 2020.
- C. Schmid. Co-organizer for Workshop on Frontiers of Video Technology, San Jose, 2017.
- J. Mairal. Member of the organizing committee of the Journées SMAI-MODE, which will take place in March 2018.

- G. Rogez. Co-organizer of the CVPR workshop on Observing and Understanding Hands in Action (HANDS 2017).

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

- C. Schmid. Area chair for ICCV 2017, NIPS 2017, ICML 2018, and ECCV 2018.
- K. Alahari. Area chair for CVPR 2018.
- J. Mairal. Area chair for ICML 2017, NIPS 2017, and ICML 2018.
- J. Verbeek. Area chair for ECCV 2018.

10.1.2.2. Reviewer

The permanent members of the team reviewed numerous papers for numerous international conferences in computer vision and machine learning: CVPR, ECCV, NIPS, ICML.

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- C. Schmid: Editor in Chief of the International Journal of Computer Vision, since 2013.
- C. Schmid: Associate editor for Foundations and Trends in Computer Graphics and Vision, since 2005.
- J. Mairal: Associate editor of the International Journal of Computer Vision (IJCV), since 2015.
- J. Mairal: Senior associate editor for IEEE Signal Processing Letters, since Feb 2015 (editor since Aug. 2014).
- J. Mairal: Associate editor of Journal of Mathematical Imaging and Vision (JMIV), since 2015.
- J. Mairal: Associate editor of the SIAM Journal of imaging science, since 2018.
- J. Verbeek: Associate editor for Image and Vision Computing Journal, since 2011.
- J. Verbeek: Associate editor for the International Journal on Computer Vision, since 2014.

10.1.3.2. Reviewer - Reviewing Activities

The permanent members of the team reviewed numerous papers for numerous international journals in computer vision (IJCV, PAMI,CVIU), machine learning (JMLR, Machine Learning). Some of them are also reviewing for journals in optimization (SIAM Journal on Optimization, Mathematical Programming), image processing (SIAM Imaging Science).

10.1.4. Invited Talks

- K. Alahari. Invited speaker at ETH Zurich Photogrammetry group retreat, Morzine, France, January 2017.
- G. Durif. Invited speaker at the “StatOmique” meeting, Paris. March 2017.
- J. Mairal. Seminar LIG. Grenoble, December 2017.
- J. Mairal. Invited speaker at Journées franco-chiliennes d’optimisation. Toulouse, July 2017.
- J. Mairal. Invited speaker at the large-scale and distributed optimization workshop, Lund, June 2017.
- J. Mairal. Invited speaker at Pattern recognition and computer vision colloquium, Prague, May 2017.
- J. Mairal. Invited speaker at the optimization and statistical learning workshop, Les Houches, April 2017.
- J. Mairal. Two seminars at Amazon Berlin. February 2017.
- G. Rogez. Invited speaker at Journées CNRS-GDR Isis, Telecom PariTech, December 2017.
- J. Verbeek. Invited talk at Aalto University, Helsinki, Finland, June 2017.
- J. Verbeek. Invited talk at ATOS, Grenoble, January 2017.

- J. Verbeek. Invited talk at Technicolor, Rennes, January 2017.
- H. Lin. Invited talk in mini-symposium. SIAM Optimization conference. May 2017.
- H. Lin. Seminar at University of Washington, June 2017.
- R. Klokov: Invited talk at ICCV workshop Learning to See from 3D Data.
- C. Schmid: Keynote speaker at ECML-PKDD 2017, Skopje, September 2017.
- C. Schmid: Keynote speaker at Gresti 2017, Juan-les-Pins, September 2017.
- C. Schmid: Invited speaker at Workshop on YouTube-8M Large-Scale Video Understanding, in conjunction with CVPR'17, July 2017.
- C. Schmid: Invited speaker at Women in Computer Vision Workshop, in conjunction with CVPR'17, July 2017.
- C. Schmid: Invited speaker at 1st Workshop on Target Re-Identification and Multi-Target Multi-Camera Tracking, in conjunction with CVPR'17, July 2017.
- C. Schmid: Invited speaker at Chalearn Looking at People Workshop, in conjunction with CVPR'17, July 2017.
- C. Schmid: Invited speaker at Frontiers of Video Technology, July 2017.
- C. Schmid: Invited speaker at Korean Conference on Computer Vision, Seoul, June 2017.
- C. Schmid: Keynote speaker at Swedish Symposium on Deep Learning, Stockholm, June 2017.o Invited speaker at Russian Summit "Machines Can See", Moscow, June 2017.
- C. Schmid: Seminar at Intel Network on Intelligent Systems, Munich, August 2017.
- C. Schmid: Seminar at Berkeley University, July 2017.
- C. Schmid: Seminar at Toyota Research Institute, July 2017.
- C. Schmid: Seminar at Google, Mountain View, July 2017.
- C. Schmid: Seminar at DeepMind, London, June 2017.
- C. Schmid: Speaker at Distinguished Seminar Series in Computing, Imperial College, London, June 2017.
- C. Schmid: Seminar at MPI, Tübingen, May 2017.
- C. Schmid: Seminar at "10 ans de l'ERC à Inria", Paris, Mars 2017.

10.1.5. Scientific Expertise

- C. Schmid: Award committee member for Oréal-UNESCO award France for Women in Sciences 2017.
- K. Alahari: reviewer for ANR.
- J. Mairal: judge for the IBM Watson AI Xprize.
- J. Mairal: reviewer for ANR.
- J. Mairal: panel member of a funding agency for the 2018 call (confidential information at the moment).
- J. Mairal: member of the PGMO best PhD prize committee in 2017.
- G. Rogez: reviewer for ANR.
- J. Verbeek: reviewer for ERC.
- J. Verbeek: reviewer for ASF.

10.1.6. Research Administration

- C. Schmid: Member of the "Comité scientifique", Inria Grenoble, since 2015.
- C. Schmid: Member of board of directors of the Computer Vision Foundation (CVF), since 2016.
- C. Schmid is member of the PAMI-TC awards committee, and the PAMI-TC executive commitee.

- J. Verbeek: Scientific correspondent national project calls, Inria Grenoble.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Doctorat: Karteek Alahari, Lecturer at the summer school on computer vision, IIIT Hyderabad, India.

Doctorat: J. Mairal, Lecturer at the SPARSA summer school, Lisbon.

Doctorat: J. Mairal, Lecturer at the CoSIP winter school, berlin.

Doctorat: J. Mairal, “Introduction to machine learning”, PSL-ITI program, 6H eq-TD, Paris.

Doctorat: J. Verbeek. Tutorial given at IPTA, Montreal.

Doctorat: J. Verbeek. Invited lecture Deep learning summer school, Univ. Côte D’Azur.

Doctorat: C. Schmid. Tutorial on action recognition at the Winter School in Computer Vision, Jerusalem, January 2017.

Master : C. Schmid, “Object recognition and computer vision”, 15H eqTD, M2, ENS Cachan, France.

Master : J. Verbeek and C. Schmid. “Machine Learning & Category Representation”, 27H eqTD, M2, Univ. Grenoble.

Master: K. Alahari, “Introduction to Discrete Optimization”, Ecole Centrale Paris, 27H eq-TD, M1, Paris, France.

Master: K. Alahari, “Discrete Inference and Learning”, Ecole Centrale Paris, M2, Paris, France.

Master: K. Alahari, “Understanding Big Visual Data”, Grenoble INP, 13.5H eq-TD, M2, Grenoble, France.

Master: K. Alahari, “Introduction to computer vision”, ENS Paris, M1, Grenoble, France.

Master: J. Mairal, “Kernel methods for statistical learning”, 27H eqTD, M2, Ecole Normale Supérieure, Cachan.

Master : J. Verbeek and J. Mairal, “Advanced Learning Models”, 27H eqTD, M2, ENSIMAG, Grenoble.

Master: A. Sablayrolles and P. Luc: Lecture at Ecole 42 on Convolutional Neural Networks as part of the Facebook AI Masterclass.

License: H. Lin, UE MAT206: Introduction à la biologie mathématique et à la dynamique des populations, 58H eqTD, L3, UGA, Grenoble.

10.2.2. Supervision

HdR: Julien Mairal, large-scale machine learning and applications, Univ. Grenoble Alpes, October 2017.

PhD: Hongzhou Lin, Generic acceleration schemes for gradient-based optimization in machine learning. November 2017. Supervision Julien Mairal and Zaid Harchaoui.

PhD: Thomas Dias-Alves, Modélisation du déséquilibre de liaison en génomique des populations par méthodes d’optimisation, Univ. Grenoble Alpes. December 2017. Supervision Julien Mairal and Michael Blum.

PhD: Mattis Paulin, Of Learning Visual Representations Robust to Invariances for Image Classification and Retrieval. March 2017. Univ. Grenoble-Alpes. Supervision Z. Harchaoui. C. Schmid, F. Perronnin, J. Mairal, and M. Douze.

PhD: Vicky Kalogeiton, Localising spatially and temporally objects and actions in videos, September 2017. Univ. Edinburgh, supervision V. Ferrari and C. Schmid.

10.2.3. Juries

K. Alahari: Raghudeep Gadde, 2017, jury member, these, Ecole des Ponts ParisTech, Paris.

K. Alahari: Lukas Neumann, 2017, rapporteur, these, Czech Technical University, Prague.

J. Mairal: Matthieu Carrière, novembre 2017, rapporteur, these, Université Paris-Saclay.

J. Mairal: Thomas Moreau, décembre 2017, rapporteur, these, Université Paris-Saclay.

J. Mairal: Olga Permiakova. novembre 2017, member of "comité de suivi de thèse", Univ. Grenoble Alpes.

J. Verbeek: Fabien Baradel. member of "comité de suivi de thèse", INSA Lyon.

J. Verbeek: Praveen Kulkarni, rapporteur, these, Univ. Caen.

J. Verbeek: Damien Fourure, Univ. Jean Monnet, Saint Etienne, France.

C. Schmid: Thibaut Durand, September 2017, examinateur, Universite Pierre et Marie Curie.

C. Schmid: Christos Georgakis, June 2017, rapport these, Imperial College, London

C. Schmid: Pedro Oliveira Pinheiro, January 2017, rapport these, EPFL.

10.3. Popularization

- J. Mairal est intervenu lors du cycle de conférences ISN, à destination des professeurs de lycée.
- J. Mairal a co-publié un article dans ERCIM news, avec Gaël Varoquaux, Bertrand Thirion et Arthur Mensch.
- C. Schmid a co-publié un article "Extraction d'informations à partir des images". Les Big Data à Découvert. Editions du CNRS, 2017

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] V. KALOGEITON. *Localizing spatially and temporally objects and actions in videos*, University of Edinburgh ; Inria Grenoble, September 2017, https://hal.inria.fr/tel-01674504

[2] J. MAIRAL. *Large-Scale Machine Learning and Applications*, UGA - Université Grenoble Alpes, October 2017, Habilitation à diriger des recherches, https://hal.inria.fr/tel-01629997

[3] M. PAULIN. *Of Learning Visual Representations Robust to Invariances for Image Classification and Retrieval*, Université Grenoble Alpes, February 2017, https://hal.inria.fr/tel-01677852

Articles in International Peer-Reviewed Journals

[4] N. CHESNEAU, K. ALAHARI, C. SCHMID. *Learning from Web Videos for Event Classification*, in "IEEE Transactions on Circuits and Systems for Video Technology", 2017 [*DOI :* 10.1109/TCSVT.2017.2764624], https://hal.inria.fr/hal-01618400

[5] R. G. CINBIS, J. VERBEEK, C. SCHMID. *Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", January 2017, vol. 39, n^{o} 1, pp. 189-203, https://arxiv.org/abs/1503.00949 [*DOI :* 10.1109/TPAMI.2016.2535231], https://hal.inria.fr/hal-01123482

[6] G. DURIF, L. MODOLO, J. MICHAELSSON, J. E. MOLD, S. LAMBERT-LACROIX, F. PICARD. *High Dimensional Classification with combined Adaptive Sparse PLS and Logistic Regression*, in "Bioinformatics", 2017, https://arxiv.org/abs/1502.05933 [*DOI :* 10.1093/BIOINFORMATICS/BTX571], https://hal.archives-ouvertes.fr/hal-01587360

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

[8] G. HU, X. PENG, Y. YANG, T. HOSPEDALES, J. VERBEEK. *Frankenstein: Learning Deep Face Representations using Small Data*, in "IEEE Transactions on Image Processing", January 2018, vol. 27, n° 1, pp. 293-303, https://arxiv.org/abs/1603.06470 [*DOI :* 10.1109/TIP.2017.2756450], https://hal.inria.fr/hal-01306168

[9] A. MENSCH, J. MAIRAL, B. THIRION, G. VAROQUAUX. *Stochastic Subsampling for Factorizing Huge Matrices*, in "IEEE Transactions on Signal Processing", January 2018, vol. 66, n° 1, pp. 113-128, https://arxiv.org/abs/1701.05363 [*DOI :* 10.1109/TSP.2017.2752697], https://hal.archives-ouvertes.fr/hal-01431618

[10] A. MISHRA, K. ALAHARI, C. JAWAHAR. *Unsupervised refinement of color and stroke features for text binarization*, in "International Journal on Document Analysis and Recognition", June 2017, vol. 20, n° 2, pp. 105–121 [*DOI :* 10.1007/S10032-017-0283-9], https://hal.inria.fr/hal-01490176

[11] M. PAULIN, J. MAIRAL, M. DOUZE, Z. HARCHAOUI, F. PERRONNIN, C. SCHMID. *Convolutional Patch Representations for Image Retrieval: an Unsupervised Approach*, in "International Journal of Computer Vision", January 2017, vol. 121, n° 1, pp. 149–168 [*DOI :* 10.1007/S11263-016-0924-3], https://hal.inria.fr/hal-01277109

[12] 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

International Conferences with Proceedings

[13] A. BIETTI, J. MAIRAL. *Invariance and Stability of Deep Convolutional Representations*, in "NIPS 2017 - 31st Conference on Advances in Neural Information Processing Systems", Los Angeles, CA, United States, December 2017, https://hal.inria.fr/hal-01630265

[14] A. BIETTI, J. MAIRAL. *Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite-Sum Structure*, in "NIPS 2017 - Advances in Neural Information Processing Systems", Long Beach, CA, United States, December 2017, pp. 1-21, https://arxiv.org/abs/1610.00970 , https://hal.inria.fr/hal-01375816

[15] N. CHESNEAU, G. ROGEZ, K. ALAHARI, C. SCHMID. *Detecting Parts for Action Localization*, in "BMVC - British Machine Vision Conference", London, United Kingdom, September 2017, https://arxiv.org/abs/1707.06005 , https://hal.inria.fr/hal-01573629

[16] N. DVORNIK, K. SHMELKOV, J. MAIRAL, C. SCHMID. *BlitzNet: A Real-Time Deep Network for Scene Understanding*, in "ICCV 2017 - International Conference on Computer Vision", Venise, Italy, October 2017, 11 p. , https://hal.archives-ouvertes.fr/hal-01573361

[17] 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

[18] V. KALOGEITON, P. WEINZAEPFEL, V. FERRARI, C. SCHMID. *Action Tubelet Detector for Spatio-Temporal Action Localization*, in "ICCV - IEEE International Conference on Computer Vision", Venice, Italy, October 2017, https://arxiv.org/abs/1705.01861 , https://hal.inria.fr/hal-01519812

[19] V. KALOGEITON, P. WEINZAEPFEL, V. FERRARI, C. SCHMID. *Joint learning of object and action detectors*, in "ICCV 2017 - IEEE International Conference on Computer Vision", Venice, Italy, October 2017, https://hal.inria.fr/hal-01575804

[20] P. LUC, N. NEVEROVA, C. COUPRIE, J. VERBEEK, Y. LECUN. *Predicting Deeper into the Future of Semantic Segmentation*, in "ICCV 2017 - International Conference on Computer Vision", Venise, Italy, October 2017, 10 p. , https://arxiv.org/abs/1703.07684 , https://hal.inria.fr/hal-01494296

[21] A. MENSCH, J. MAIRAL, D. BZDOK, B. THIRION, G. VAROQUAUX. *Learning Neural Representations of Human Cognition across Many fMRI Studies*, in "Neural Information Processing Systems", Long Beach, United States, December 2017, https://arxiv.org/abs/1710.11438 , https://hal.archives-ouvertes.fr/hal-01626823

[22] M. PEDERSOLI, T. LUCAS, C. SCHMID, J. VERBEEK. *Areas of Attention for Image Captioning*, in "ICCV - International Conference on Computer Vision", Venice, Italy, October 2017, https://hal.inria.fr/hal-01428963

[23] 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

[24] G. ROGEZ, P. WEINZAEPFEL, C. SCHMID. *LCR-Net: Localization-Classification-Regression for Human Pose*, in "CVPR 2017 - IEEE Conference on Computer Vision & Pattern Recognition", Honolulu, United States, June 2017, https://hal.inria.fr/hal-01505085

[25] K. SHMELKOV, C. SCHMID, K. ALAHARI. *Incremental Learning of Object Detectors without Catastrophic Forgetting*, in "ICCV - IEEE International Conference on Computer Vision", Venice, Italy, October 2017, https://hal.inria.fr/hal-01573623

[26] P. TOKMAKOV, K. ALAHARI, C. SCHMID. *Learning Motion Patterns in Videos*, in "CVPR - IEEE Conference on Computer Vision & Pattern Recognition", Honolulu, United States, July 2017, https://arxiv.org/abs/1612.07217 , https://hal.archives-ouvertes.fr/hal-01427480

[27] P. TOKMAKOV, K. ALAHARI, C. SCHMID. *Learning Video Object Segmentation with Visual Memory*, in "ICCV - IEEE International Conference on Computer Vision", Venice, Italy, October 2017, https://arxiv.org/abs/1704.05737 , https://hal.archives-ouvertes.fr/hal-01511145

[28] 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

Other Publications

[29] A. BIETTI, J. MAIRAL. *Group Invariance, Stability to Deformations, and Complexity of Deep Convolutional Representations*, November 2017, https://arxiv.org/abs/1706.03078 - working paper or preprint, https://hal.inria.fr/hal-01536004

[30] D. CHEN, L. JACOB, J. MAIRAL. *Predicting Transcription Factor Binding Sites with Convolutional Kernel Networks*, November 2017, working paper or preprint, https://hal.inria.fr/hal-01632912

[31] T. DIAS-ALVES, J. MAIRAL, M. BLUM. *Loter: A software package to infer local ancestry for a wide range of species*, November 2017, working paper or preprint [*DOI :* 10.1101/213728], https://hal.inria.fr/hal-01630228

[32] G. DURIF, L. MODOLO, J. E. MOLD, S. LAMBERT-LACROIX, F. PICARD. *Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis*, November 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01649275

[33] T. LUCAS, J. VERBEEK. *Auxiliary Guided Autoregressive Variational Autoencoders*, November 2017, working paper or preprint, https://hal.inria.fr/hal-01652881

[34] H. LIN, J. MAIRAL, Z. HARCHAOUI. *A Generic Quasi-Newton Algorithm for Faster Gradient-Based Optimization*, April 2017, https://arxiv.org/abs/1610.00960 - working paper or preprint, https://hal.inria.fr/hal-01376079

[35] H. LIN, J. MAIRAL, Z. HARCHAOUI. *Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice*, December 2017, working paper or preprint, https://hal.inria.fr/hal-01664934

[36] C. PAQUETTE, H. LIN, D. DRUSVYATSKIY, J. MAIRAL, Z. HARCHAOUI. *Catalyst Acceleration for Gradient-Based Non-Convex Optimization*, June 2017, working paper or preprint, https://hal.inria.fr/hal-01536017

[37] P. TOKMAKOV, C. SCHMID, K. ALAHARI. *Learning to Segment Moving Objects*, December 2017, https://arxiv.org/abs/1712.01127 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01653720

[38] K. TOMBRE, L. QUAN, R. HORAUD, P. GROS, C. SCHMID, P. STURM. *In Memoriam Roger Mohr*, Société Informatique de France, September 2017, pp. 91-98, Article qui rappelle la carrière scientifique de Roger Mohr, https://hal.inria.fr/hal-01598085

[39] N. VERMA, E. BOYER, J. VERBEEK. *Dynamic Filters in Graph Convolutional Networks*, June 2017, https://arxiv.org/abs/1706.05206 - working paper or preprint, https://hal.inria.fr/hal-01540389

[40] P. WEINZAEPFEL, X. MARTIN, C. SCHMID. *Human Action Localization with Sparse Spatial Supervision*, May 2017, https://arxiv.org/abs/1605.05197 - working paper or preprint, https://hal.inria.fr/hal-01317558