Activity Report 2016

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

1. Members **1**
2. Overall Objectives **2**
3. Research Program **3**
3.1. Designing and learning structured models 3
3.2. Learning of visual models from minimal supervision 4
3.3. Large-scale learning and optimization 6
3.4. Datasets and evaluation 7
4. Application Domains **8**
5. Highlights of the Year **9**
6. New Software and Platforms **9**
6.1. CoNFab: COnvolutional Neural FABric 9
6.2. Modl 9
6.3. M-CNN: Weakly-Supervised Semantic Segmentation using Motion Cues 9
6.4. DALY: Daily Action Localization in Youtube 10
6.5. GUN-71 10
6.6. Synthetic human 3D pose dataset 10
7. New Results **10**
7.1. Visual recognition in images 10
7.1.1. Convolutional Neural Fabrics 10
7.1.2. Heterogeneous Face Recognition with CNNs 10
7.1.3. Mocap-guided Data Augmentation for 3D Pose Estimation in the Wild 12
7.1.4. End-to-End Kernel Learning with Supervised Convolutional Kernel Networks 12
7.1.5. Semantic segmentation using Adversarial Networks 12
7.1.6. Enhancing Energy Minimization Framework for Scene Text Recognition with Top-Down Cues 12
7.1.7. Local Convolutional Features with Unsupervised Training for Image Retrieval 14
7.2. Visual recognition in videos 16
7.2.1. Towards Weakly-Supervised Action Localization 16
7.2.2. The DALY dataset 16
7.2.3. Weakly-Supervised Semantic Segmentation using Motion Cues 17
7.2.4. Multi-region two-stream R-CNN for action detection 17
7.2.5. Analysing domain shift factors between videos and images for object detection 18
7.3. Large-scale statistical learning 19
7.3.1. Dictionary Learning for Massive Matrix Factorization 19
7.3.2. Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite-Sum Structure 19
7.3.3. QuickeNing: A Generic Quasi-Newton Algorithm for Faster Gradient-Based Optimization 20
7.3.4. Dictionary Learning from Phaseless Measurements 20
8. Bilateral Contracts and Grants with Industry **21**
8.1. MSR-Inria joint lab: scientific image and video mining 21
8.2. MSR-Inria joint lab: structured large-scale machine learning 21
8.3. Amazon 21
8.4. Google 22
8.5. Facebook 22
8.6. MBDA 22
8.7. Xerox Research Center Europe 22
9. Partnerships and Cooperations **22**
9.1. Regional Initiatives 22

9.2. National Initiatives 23
9.2.1. ANR Project Physionomie 23
9.2.2. ANR Project Macaron 23
9.2.3. ANR Project DeepInFrance 23
9.3. European Initiatives 23
9.3.1.1. ERC Advanced grant Allegro 23
9.3.1.2. EU Marie Curie project: Egovision4health 24
9.4. International Initiatives 24
9.4.1. Inria Associate Teams Not Involved in an Inria International Labs 24
9.4.2. Inria International Partners 24
9.4.3. Participation in Other International Programs 25
9.5. International Research Visitors 25
10. Dissemination **25**
10.1. Promoting Scientific Activities 25
10.1.1. Scientific Events Organisation 25
10.1.2. Scientific Events Selection 25
10.1.2.1. Member of the Conference Program Committees 25
10.1.2.2. Reviewer 25
10.1.3. Journal 26
10.1.3.1. Member of the Editorial Boards 26
10.1.3.2. Reviewer - Reviewing Activities 26
10.1.4. Invited Talks 26
10.1.5. Scientific Expertise 27
10.1.6. Research Administration 27
10.2. Teaching - Supervision - Juries 27
10.2.1. Teaching 27
10.2.2. Supervision 27
10.2.3. Juries 28
11. Bibliography **28**

Project-Team THOTH

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

Keywords:

Computer Science and Digital Science:

3.4. - Machine learning and statistics
5.4. - Computer vision

Other Research Topics and Application Domains:

5.6. - Robotic systems
5.8. - Learning and training
7.2. - Smart travel
8.4. - Security and personal assistance
8.5. - Smart society

1. Members

Research Scientists

Cordelia Schmid [Team leader, Inria, Senior Researcher, HDR]
Karteek Alahari [Inria, Researcher]
Julien Mairal [Inria, Researcher, "en détachement du Corps des Mines"]
Jakob Verbeek [Inria, Researcher, HDR]
Grégory Rogez [Inria, Starting Research position, funded by FP7 Marie Curie IOF - Egovision4health, from Jul 2015 to Jun 2016 and by ERC Allegro from July 2016 to June 2017]

Engineers

Julien Bardonnet [Inria, funded by MBDA, until Apr 2016]
Xavier Martin [Inria, funded by ERC Allegro]

PhD Students

Alberto Bietti [Univ. Grenoble, funded by MSR-Inria joint centre, from May 2016 to Sep 2019]
Guilhem Cheron [Ens, funded by MSR/Inria, from Oct 2014 until Oct 2017, co-supervision with I. Laptev]
Nicolas Chesneau [Univ. Grenoble, funded by ERC Allegro, from Jul 2014 until Sep 2017]
Thomas Dias-Alves [Univ. Grenoble, co-supervised with M. Blum (TIMC laboratory), from Oct 2014 to Sep 2017]
Mikita Dvornik [Univ. Grenoble, funded by ERC Allegro and ANR Macaron, from Feb 2016 to Sep 2019]
Maha Elbayad [Univ. Grenoble, funded by Persyval DeCoRe project, from Oct 2016 until Sep 2019]
Yang Hua [Univ. Grenoble, funded by MSR/Inria joint lab, from Jan 2013 until June 2016]
Vicky Kalogeiton [Univ. Edinburgh, funded by European Research Council, co-supervision with V. Ferrari, from Sep 2013 until July 2017]
Hongzhou Lin [Univ. Grenoble, funded by Université Grenoble Alpes, from Apr 2014 until Sep 2017]
Pauline Luc [Univ. Grenoble, funded by Facebook, from Jan 2016 to Dec 2018]
Thomas Lucas [Univ. Grenoble, funded by Université Grenoble Alpes, from Feb 2016 to Sep 2019]
Mattis Paulin [Univ. Grenoble, funded by DGA, from Apr 2013 until Apr 2016]
Federico Pierucci [Univ. Grenoble, funded by Université Grenoble Alpes, from Jan 2012 until March 2016]
Shreyas Saxena [Univ. Grenoble, ANR PHYSIONOMIE project, from Feb 2013 until Dec 2016]
Konstantin Shmelkov [Univ. Grenoble, funded by ERC Allegro, from Oct 2015 until Oct 2018]
Vladyslav Sydorov [Univ. Grenoble, funded by ERC Allegro and CEFIPRA, from Feb 2015 until Oct 2019]
Pavel Tokmakov [Univ. Grenoble, funded by ERC Allegro, from Sep 2014 until Sep 2017]

Philippe Weinzaepfel [Univ. Grenoble, funded by Université Grenoble Alpes, from Nov 2012 until Sep 2016]
Daan Wynen [Univ. Grenoble, funded by ERC Allegro and ANR Macaron, from Oct 2015 to Sep 2018]

Post-Doctoral Fellows

Guosheng Hu [Inria, funded by ANR Physionomie project, until May 2016]
Henrique Morimitsu [Inria, funded by ERC Allegro, from Mar 2016]
Marco Pedersoli [Inria, funded by ERC Allegro and MBDA project, until Nov 2016]
Xiaojiang Peng [Inria, funded by ERC Allegro, until Mar 2016]

Administrative Assistant

Nathalie Gillot [Inria]

Others

Dexiong Chen [Ecole Polytechnique, Intern, from Mar 2016 until August 2016]
Erwan Le Roux [ENSIMAG, Intern, from Feb 2016 to June 2016]
Valentin Thomas [ENS Cachan, Intern, from Apr 2016 until Oct 2016]

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 transforming 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 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

[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.

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.

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

- Cordelia Schmid received the Humboldt Research Award, granted by the Alexander von Humboldt Foundation.
- Cordelia Schmid was awarded the Longuet-Higgins Prize at CVPR 2016 for the paper co-authored with Svetlana Lazebnik (University of Illinois at Urbana-Champaign) and Jean Ponce (ENS Paris/Inria) entitled "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories".
- Cordelia Schmid was awarded the Inria - Académie des Sciences Grand Prize 2016.
- Thoth is one of the recipients of a hardware donation in the Facebook AI Research Partnership Program.
- Julien Mairal was awarded one of the ERC starting grants 2016.

6. New Software and Platforms

6.1. CoNFab: COnvolutional Neural FABric

Participants: Shreyas Saxena, Jakob Verbeek.

Despite the success of convolutional neural networks, selecting the optimal architecture for a given task remains an open problem. Instead of aiming to select a single optimal architecture, we propose Convolutional Neural Fabrics [20] that embed an exponentially large class of CNN architectures. The fabric consists of a 3D trellis that connects response maps at different layers, scales, and channels with a sparse homogeneous local connectivity pattern. The only hyper-parameters of the model (nr. of channels and layers) are not critical for performance. While individual CNN architectures can be recovered as paths in the trellis, the trellis can in addition ensemble all embedded architectures together, sharing their weights where their paths overlap. By the non-cyclic property of the trellis, its parameters can be efficiently learned using methods based on error back-propagation. The trellis parameters can be learned using standard methods based on back-propagation, at a cost that scales linearly in the fabric size. This software implements Convolutional Neural Fabrics by means of wrappers on top of the Caffe library to specify and learn such models.

6.2. Modl

Participants: Julien Mairal, Arthur Mensch [Parietal], Gael Varoquaux [Parietal], Bertrand Thirion [Parietal].

Modl is a new Python library written by Arthur Mensch for factorizing huge matrices. It implements the method presented in [25], [17], which targets matrices of several terabytes that do not fit into the main computer's memory.

6.3. M-CNN: Weakly-Supervised Semantic Segmentation using Motion Cues

Participants: Pavel Tokmakov, Cordelia Schmid, Karteek Alahari.

This is a public implementation of the method described in [23]. It includes a framework for integrating motion cues into training a deep network for weakly-supervised semantic segmentation, code for data preprocessing and trained models that correspond to the results reported in the paper. Our code is built on top of DeepLab https://bitbucket.org/aquariusjay/deeplab-public-ver2 extension of the Caffe deep learning framework http://caffe.berkeleyvision.org.

6.4. DALY: Daily Action Localization in Youtube

Participants: Philippe Weinzapfael, Xavier Martin, Cordelia Schmid.

DALY is a video dataset with spatial and temporal annotation of 10 everyday human actions in 31 hours of Youtube videos, which allows to train and benchmark methods for action recognition and localization in videos. It is available at http://thoth.inrialpes.fr/daly/. We developed the dataset jointly with a new action localization technique. Both are described in [33].

6.5. GUN-71

Participant: Gregory Rogez.

This dataset consist of 12,000 RGB-D images of object manipulation scenes (captured from a chest-mounted camera) that were labeled with one of 71 fine-grained grasps. We considered 28 objects per grasp, resulting in a total of 1988 different hand-object configurations with 5-6 views for each. The data were captured with 8 different subjects (4 males and 4 females) in 5 different houses, see http://www.gregrogez.net/research/egovision4health/gun-71/.

6.6. Synthetic human 3D pose dataset

Participants: Gregory Rogez, Cordelia Schmid.

Participants: Gregory Rogez, Cordelia Schmid This large-scale dataset consists of 2,000,000 artificial RGB images of humans and associated 2D and 3D pose annotations. This dataset was generated using the image-based rendering algorithm presented in [19] and has been used to train state-of-the-art Convolutional Neural Networks (CNN) for in-the-wild 3D human pose estimation, see http://www.gregrogez.net/research/human-pose-data-synthesis-for-cnn/.

7. New Results

7.1. Visual recognition in images

7.1.1. Convolutional Neural Fabrics

Participants: Shreyas Saxena, Jakob Verbeek.

Despite the success of CNNs, selecting the optimal architecture for a given task remains an open problem. Instead of aiming to select a single optimal architecture, in this work [20], we propose a "fabric" that embeds an exponentially large number of architectures. See 1 for a schematic illustration of how fabrics embed different architectures. The fabric consists of a 3D trellis that connects response maps at different layers, scales, and channels with a sparse homogeneous local connectivity pattern. The only hyper-parameters of a fabric are the number of channels and layers. While individual architectures can be recovered as paths, the fabric can in addition ensemble all embedded architectures together, sharing their weights where their paths overlap. Parameters can be learned using standard methods based on back-propagation, at a cost that scales linearly in the fabric size. We present benchmark results competitive with the state of the art for image classification on MNIST and CIFAR10, and for semantic segmentation on the Part Labels dataset.

7.1.2. Heterogeneous Face Recognition with CNNs

Participants: Shreyas Saxena, Jakob Verbeek.

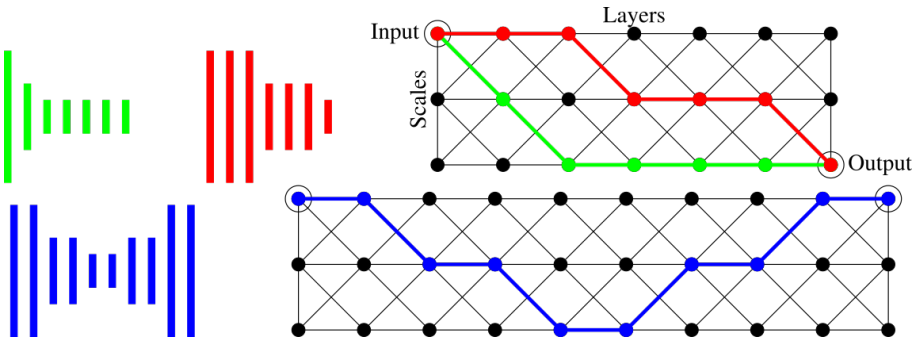

Figure 1. Fabrics embedding two seven-layer CNNs (red, green) and a ten-layer deconvolutional network (blue). Feature map size of the CNN layers are given by height. Fabric nodes receiving input and producing output are encircled. All edges are oriented to the right, down in the first layer, and towards the output in the last layer. The channel dimension of the 3D fabric is omitted for clarity.

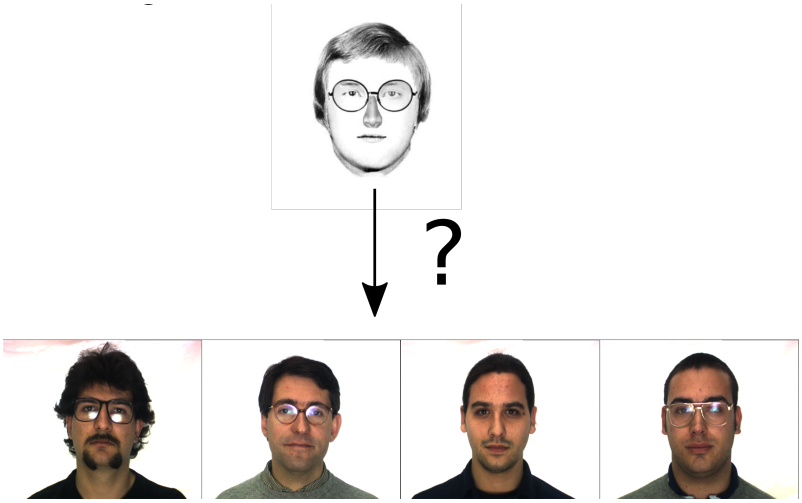

Figure 2. Schematic illustration for the task of heterogenous face recognition. The goal is to find the identity of the probe image (shown as a sketch) among one of the identities from the gallery set (shown in the bottom row). In contrast to standard face recognition, the probe and the gallery set do not share the same modality. In the illustartion, the probe image is a sketch and the galley images are normal visible spectrum images.

Heterogeneous face recognition aims to recognize faces across different sensor modalities, see 2 for a schematic illustration. Typically, gallery images are normal visible spectrum images, and probe images are infrared images or sketches. Recently significant improvements in visible spectrum face recognition have been obtained by CNNs learned from very large training datasets. In this paper [21], we are interested in the question to what extent the features from a CNN pre-trained on visible spectrum face images can be used to perform heterogeneous face recognition. We explore different metric learning strategies to reduce the discrepancies between the different modalities. Experimental results show that we can use CNNs trained on visible spectrum images to obtain results that are on par or improve over the state-of-the-art for heterogeneous recognition with near-infrared images and sketches.

7.1.3. Mocap-guided Data Augmentation for 3D Pose Estimation in the Wild

Participants: Grégory Rogez, Cordelia Schmid.

In this paper [19], we address the problem of 3D human pose estimation in the wild. A significant challenge is the lack of training data, i.e., 2D images of humans annotated with 3D poses. Such data is necessary to train state-of-the-art CNN architectures. Here, we propose a solution to generate a large set of photorealistic synthetic images of humans with 3D pose annotations. We introduce an image-based synthesis engine that artificially augments a dataset of real images with 2D human pose annotations using 3D Motion Capture (MoCap) data. Given a candidate 3D pose our algorithm selects for each joint an image whose 2D pose locally matches the projected 3D pose. The selected images are then combined to generate a new synthetic image by stitching local image patches in a kinematically constrained manner. See examples in Figure 3. The resulting images are used to train an end-to-end CNN for full-body 3D pose estimation. We cluster the training data into a large number of pose classes and tackle pose estimation as a K-way classification problem. Such an approach is viable only with large training sets such as ours. Our method outperforms the state of the art in terms of 3D pose estimation in controlled environments (Human3.6M) and shows promising results for in-the-wild images (LSP). This demonstrates that CNNs trained on artificial images generalize well to real images.

7.1.4. End-to-End Kernel Learning with Supervised Convolutional Kernel Networks

Participant: Julien Mairal.

In [16], we introduce 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). We show that our method achieves reasonably competitive performance for image classification on some standard " deep learning " datasets such as CIFAR-10 and SVHN, and also for image super-resolution, demonstrating the applicability of our approach to a large variety of image-related tasks. The model is illustrated in Figure 4.

7.1.5. Semantic segmentation using Adversarial Networks

Participants: Pauline Luc, Camille Couprie [Facebook], Soumith Chintala [Facebook], Jakob Verbeek.

Adversarial training has been shown to produce state of the art results for generative image modeling. In [24], we propose an adversarial training approach to train semantic segmentation models. We train a convolutional semantic segmentation network along with an adversarial network that discriminates segmentation maps coming either from the ground truth or from the segmentation network, as shown in Figure 5. The motivation for our approach is that it can detect and correct higher-order inconsistencies between ground truth segmentation maps and the ones produced by the segmentation net. Our experiments show that our adversarial training approach leads to improved accuracy on the Stanford Background and PASCAL VOC 2012 datasets.

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

Participants: Anand Mishra [IIIT Hyderabad], Karteek Alahari, C. v. Jawahar [IIIT Hyderabad].

Figure 3. Given a candidate 3D pose, our algorithm selects for each joint an image whose annotated 2D pose locally matches the projected 3D pose. The selected images are then combined to generate a new synthetic image by stitching local image patches in a kinematically constrained manner. We show 6 examples corresponding to the same 3D pose observed from 6 different camera viewpoints.

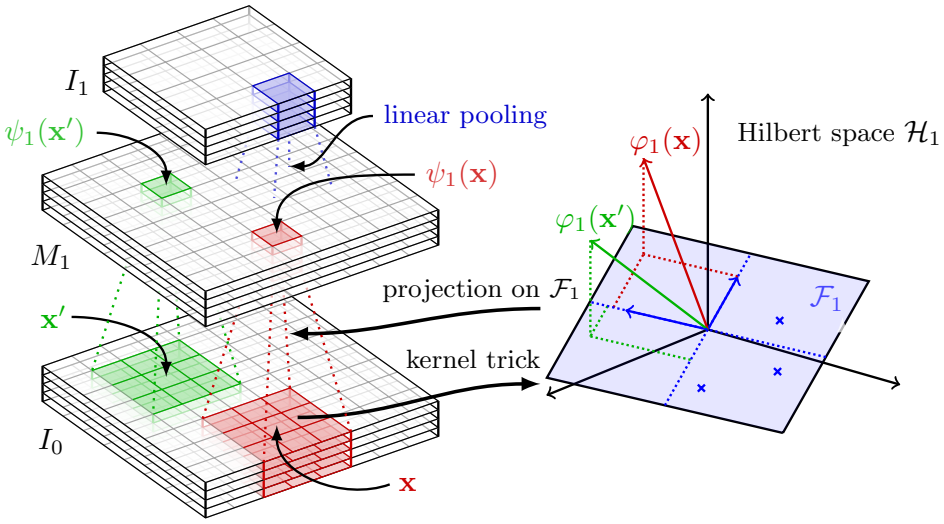

Figure 4. Our variant of convolutional kernel networks, illustrated between layers 0 and 1.

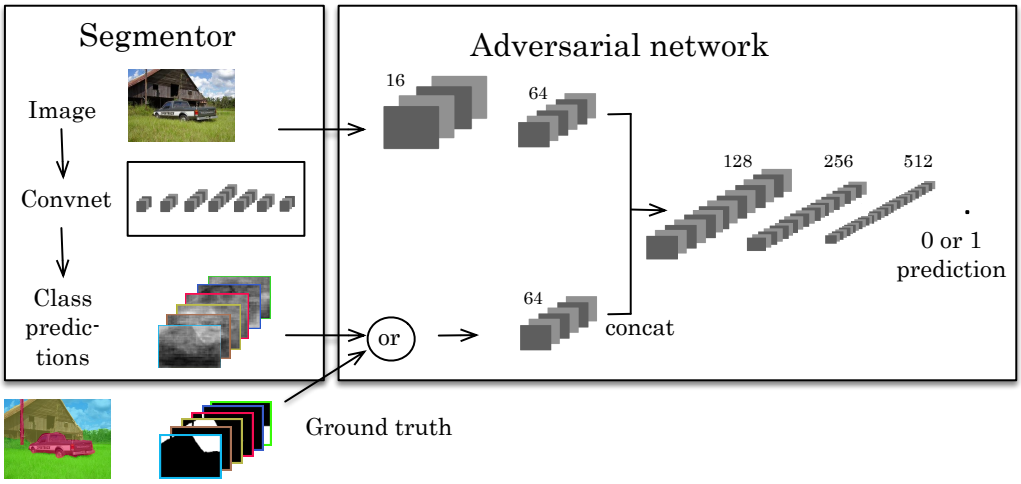

Figure 5. We use adversarial training to simultaneously learn a segmentation model (left) and a high order loss term to train it, given by the adversarial network (right). This encourages the segmentation model to output plausible segmentations, by enforcing forms of high order consistencies that are learned rather than manually designed.

Recognizing scene text, i.e., text in images such as the one in Figure 6, is a challenging problem, even more so than the recognition of scanned documents. This problem has gained significant attention from the computer vision community in recent years, and several methods based on energy minimization frameworks and deep learning approaches have been proposed. In our work presented in [8], we focus on the energy minimization framework and propose a model that exploits both bottom-up and top-down cues for recognizing cropped words extracted from street images. The bottom-up cues are derived from individual character detections from an image. We build a conditional random field model on these detections to jointly model the strength of the detections and the interactions between them. These interactions are top-down cues obtained from a lexicon-based prior, i.e., language statistics. The optimal word represented by the text image is obtained by minimizing the energy function corresponding to the random field model. We evaluate our proposed algorithm extensively on a number of cropped scene text benchmark datasets, namely Street View Text, ICDAR 2003, 2011 and 2013 datasets, and IIIT 5K-word, and show better performance than comparable methods. We perform a rigorous analysis of all the steps in our approach and analyze the results. We also show that state-of-the-art convolutional neural network features can be integrated in our framework to further improve the recognition performance.

7.1.7. Local Convolutional Features with Unsupervised Training for Image Retrieval

Participants: Mattis Paulin, Matthijs Douze [Facebook], Zaid Harchaoui [University of Washington], Julien Mairal, Florent Perronnin [Xerox], Cordelia Schmid.

Patch-level descriptors underlie several important computer vision tasks, such as stereo-matching or content-based image retrieval. We introduce a deep convolutional architecture that yields patch-level descriptors, as an alternative to the popular SIFT descriptor for image retrieval. The proposed family of descriptors, called Patch-CKN[9], adapt the recently introduced Convolutional Kernel Network (CKN), an unsupervised framework to learn convolutional architectures. We present a comparison framework to benchmark current deep convolutional approaches along with Patch-CKN for both patch and image retrieval (see Fig. 7 for our pipeline), including our novel "RomePatches" dataset. Patch-CKN descriptors yield competitive results compared to supervised CNNs alternatives on patch and image retrieval.

Figure 6. A typical street scene image taken from Google Street View. It contains very prominent sign boards with text on the building and its windows. It also contains objects such as car, person, tree, and regions such as road, sky. Many scene understanding methods recognize these objects and regions in the image successfully, but overlook the text on the sign board, which contains rich, useful information. The goal of our work [8] is to address this gap in understanding scenes.

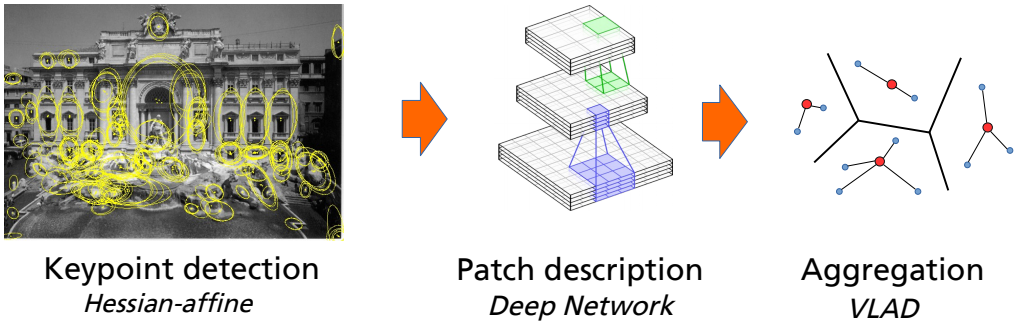

Figure 7. Image retrieval pipeline. Interest points are extracted with the Hessian-affine detector (left), encoded in descriptor space using convolutional features (middle), and aggregated into a compact representation using VLAD-pooling (right).

7.2. Visual recognition in videos

7.2.1. Towards Weakly-Supervised Action Localization

Participants: Philippe Weinzaepfel, Xavier Martin, Cordelia Schmid.

In this paper [33], we present a novel approach for weakly-supervised action localization, i.e., that does not require per-frame spatial annotations for training. We first introduce an effective method for extracting human tubes by combining a state-of-the-art human detector with a tracking-by-detection approach. Our tube extraction leverages the large amount of annotated humans available today and outperforms the state of the art by an order of magnitude: with less than 5 tubes per video, we obtain a recall of 95% on the UCF-Sports and J-HMDB datasets. Given these human tubes, we perform weakly-supervised selection based on multi-fold Multiple Instance Learning (MIL) with improved dense trajectories and achieve excellent results. Figure 8 summarizes the approach. We obtain a mAP of 84% on UCF-Sports, 54% on J-HMDB and 45% on UCF-101, which outperforms the state of the art for weakly-supervised action localization and is close to the performance of the best fully-supervised approaches. The second contribution of this paper is a new realistic dataset for action localization, named DALY (Daily Action Localization in YouTube). It contains high quality temporal and spatial annotations for 10 actions in 31 hours of videos (3.3M frames), which is an order of magnitude larger than standard action localization datasets. On the DALY dataset, our tubes have a spatial recall of 82%, but the detection task is extremely challenging, we obtain 10.8% mAP.

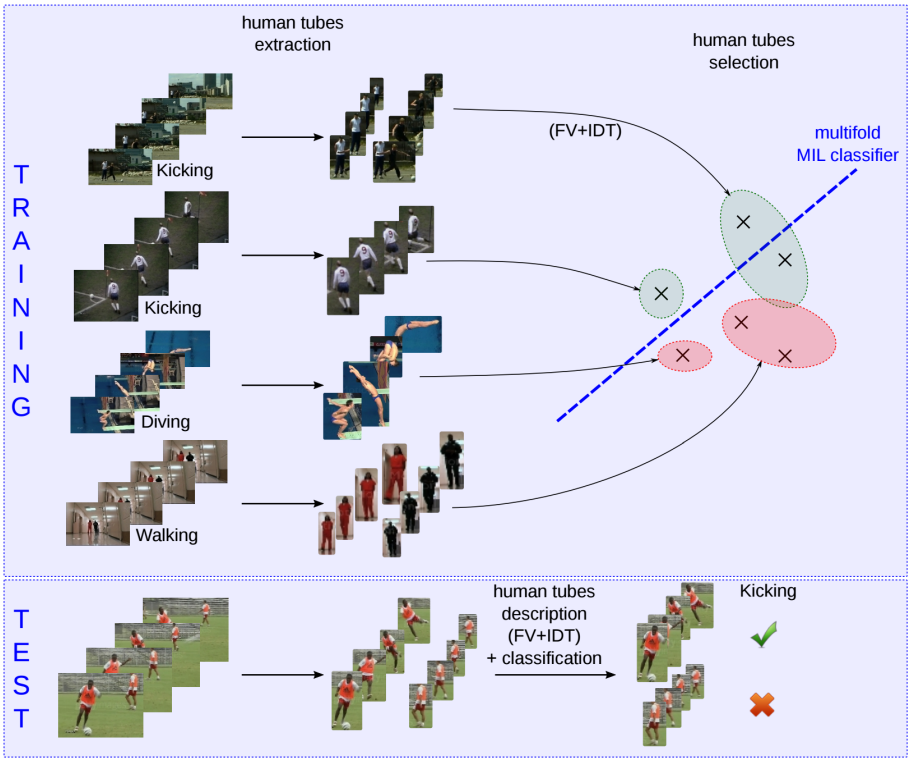

Figure 8. Overview of our approach for action localization without spatial supervision.

7.2.2. The DALY dataset

Participants: Philippe Weinzaepfel, Xavier Martin, Cordelia Schmid.

We introduce a new action localization dataset named DALY (Daily Action Localization in YouTube). DALY consists of more than 31 hours of videos (3.3M frames) from YouTube with 10 realistic daily actions, see Figure 9, and 3.6k spatio-temporal instances. Annotations consist in the start and end time of each action instance, with high-quality spatial annotation for a sparse subset of frames. The task is to localize relatively short actions (8 seconds in average) in long untrimmed videos (3min 45 in average). Furthermore, it includes videos with multiple humans performing actions simultaneously. It overcomes the limitations of existing benchmarks that are limited to trimmed or almosttrimmed videos with specific action types, e.g. sports only, showing in most cases one human per video.

Figure 9. Overview of our approach for action localization without spatial supervision.

7.2.3. Weakly-Supervised Semantic Segmentation using Motion Cues

Participants: Pavel Tokmakov, Karteek Alahari, Cordelia Schmid.

Fully convolutional neural networks (FCNNs) trained on a large number of images with strong pixel-level annotations have become the new state of the art for the semantic segmentation task. While there have been recent attempts to learn FCNNs from image-level weak annotations, they need additional constraints, such as the size of an object, to obtain reasonable performance. To address this issue, in [23] we present motion-CNN (M-CNN), a novel FCNN framework which incorporates motion cues and is learned from video-level weak annotations. Our learning scheme to train the network uses motion segments as soft constraints, thereby handling noisy motion information, as shown in Figure 10. When trained on weakly-annotated videos, our method outperforms the state-of-the-art EM-Adapt approach on the PASCAL VOC 2012 image segmentation benchmark. We also demonstrate that the performance of M-CNN learned with 150 weak video annotations is on par with state-of-the-art weakly-supervised methods trained with thousands of images. Finally, M-CNN substantially outperforms recent approaches in a related task of video co-localization on the YouTube-Objects dataset.

7.2.4. Multi-region two-stream R-CNN for action detection

Participants: Xiaojiang Peng, Cordelia Schmid.

This work [18] introduces a multi-region two-stream R-CNN model for action detection, see Figure 11. It starts from frame-level action detection based on faster R-CNN and makes three contributions. The first one is the introduction of a motion region proposal network (RPN) complementary to a standard appearance RPN. The second is the stacking of optical flow over several frames, which significantly improves frame-level action detection. The third is the addition of a multi-region scheme to the faster R-CNN model, which adds complementary information on body parts. Frame-level detections are linked with the Viterbi algorithm, and action are temporally localized with the maximum subarray method. Experimental results on the UCF-Sports,

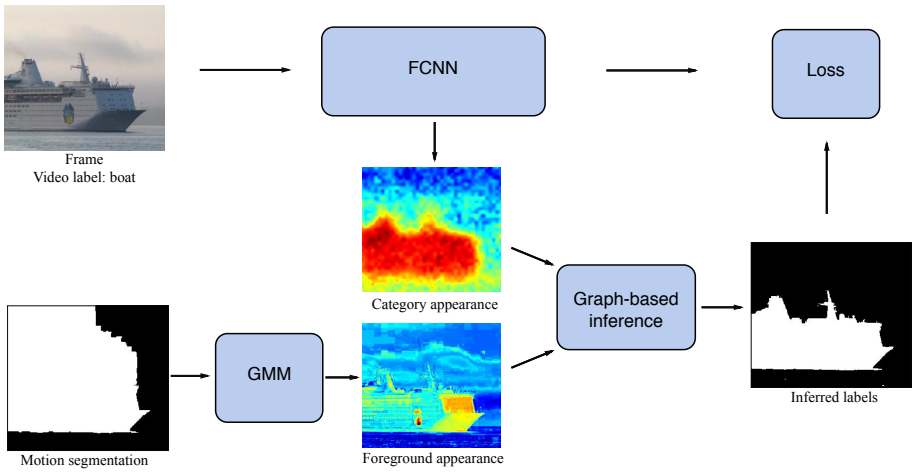

Figure 10. Overview of our M-CNN framework, where we show only one frame from a video example for clarity. The soft potentials (foreground appearance) computed from motion segmentation and the FCNN predictions (category appearance) jointly determine the latent segmentation (inferred labels) to compute the loss, and thus the network update.

J-HMDB and UCF101 action detection datasets show that the approach outperforms the state of the art with a significant margin in both frame-mAP and video-mAP.

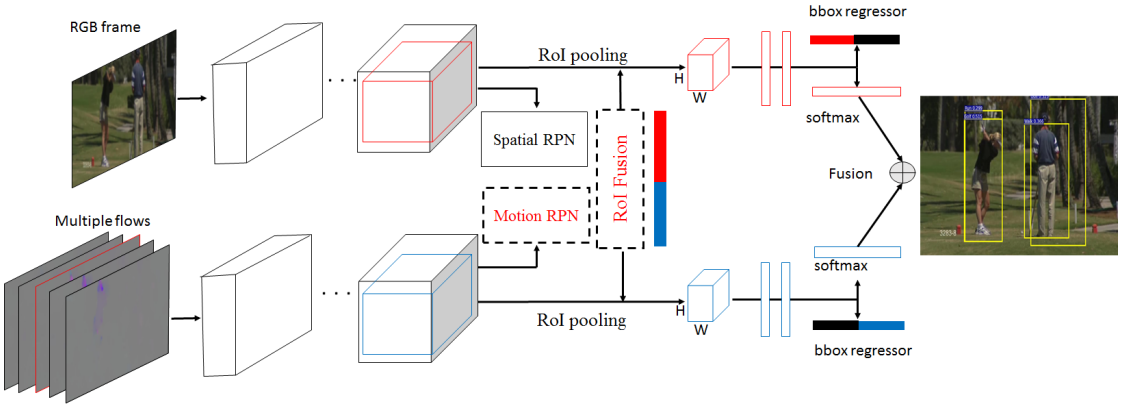

Figure 11. Two-stream faster R-CNN for spatio-temporal action detection.

7.2.5. Analysing domain shift factors between videos and images for object detection

Participants: Vicky Kalogeiton, Vittorio Ferrari [Univ. Edinburgh], Cordelia Schmid.

Object detection is one of the most important challenges in computer vision. Object detectors are usually trained on bounding-boxes from still images. Recently, video has been used as an alternative source of data. Yet, for a given test domain (image or video), the performance of the detector depends on the domain it was

trained on. In this paper [7], we examine the reasons behind this performance gap. We define and evaluate different domain shift factors (see Figure 12): spatial location accuracy, appearance diversity, image quality and aspect distribution. We examine the impact of these factors by comparing performance before and after factoring them out. The results show that all four factors affect the performance of the detectors and their combined effect explains nearly the whole performance gap.

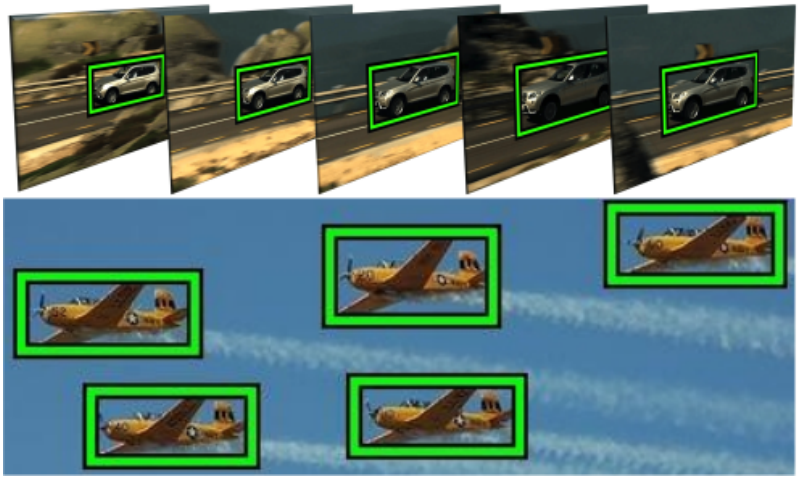

Figure 12. Example of apperance diversity domain shift factor. (top row): Frames in the same shot that contain near identical samples of an object. (bottom row): Example of near identical samples in the same image.

7.3. Large-scale statistical learning

7.3.1. Dictionary Learning for Massive Matrix Factorization

Participants: Julien Mairal, Arthur Mensch [Parietal], Gael Varoquaux [Parietal], Bertrand Thirion [Parietal].

Sparse matrix factorization is a popular tool to obtain interpretable data decompositions, which are also effective to perform data completion or denoising. Its applicability to large datasets has been addressed with online and randomized methods, that reduce the complexity in one of the matrix dimension, but not in both of them. In [25], [17], we tackle very large matrices in both dimensions. We propose a new factorization method that scales gracefully to terabyte-scale datasets. Those could not be processed by previous algorithms in a reasonable amount of time. We demonstrate the efficiency of our approach on massive functional Magnetic Resonance Imaging (fMRI) data, and on matrix completion problems for recommender systems, where we obtain significant speed-ups compared to state-of-the art coordinate descent methods. The main principle of the method is illustrated in Figure 13.

7.3.2. 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. However, in the context of empirical risk minimization, it is often helpful to augment the training set by considering random perturbations of input examples. In this case, the objective is no longer a finite sum, and the main candidate for optimization is the stochastic gradient descent method (SGD). In this paper [26], we introduce a variance reduction approach for this setting when the objective is strongly convex.

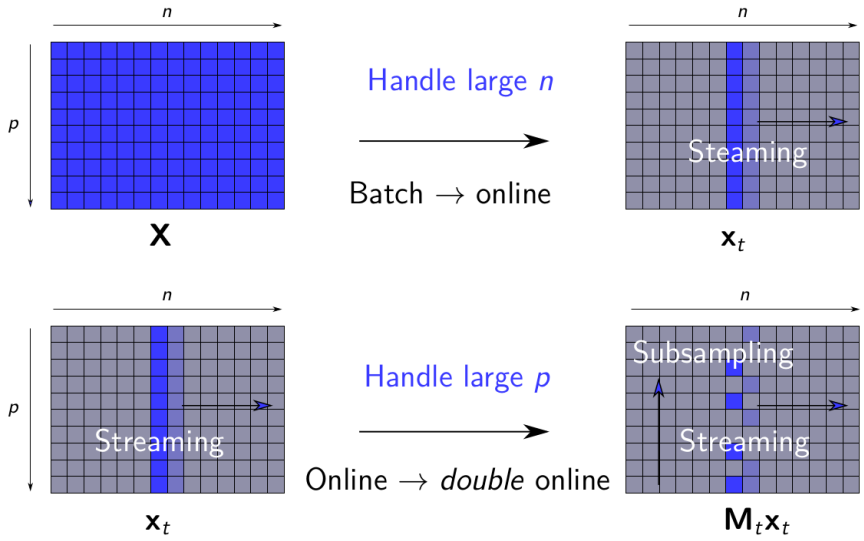

Figure 13. Illustration of the matrix factorization algorithm, which streams columns in one dimension while subsampling them.

After an initial linearly convergent phase, the algorithm achieves a $O(1/t)$ convergence rate in expectation like SGD, but with a constant factor that is typically much smaller, depending on the variance of gradient estimates due to perturbations on a single example.

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

Participants: Hongzhou Lin, Julien Mairal, Zaid Harchaoui [University of Washington].

In this paper [28], we propose an approach to accelerate gradient-based optimization algorithms by giving them the ability to exploit curvature information using quasi-Newton update rules. The proposed scheme, called QuickeNing, is generic and can be applied to a large class of first-order methods such as incremental and block-coordinate algorithms; 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; with no line-search, it is also simple to use and to implement. Besides, it enjoys a worst-case linear convergence rate for strongly convex problems. We present experimental results, see Figure 14, where QuickeNing gives significant improvements over competing methods for solving large-scale high-dimensional machine learning problems.

7.3.4. Dictionary Learning from Phaseless Measurements

Participants: Julien Mairal, Yonina Eldar [Technion], Andreas Tillmann [TU Darmstadt].

In [22], [12], we propose a new algorithm to learn a dictionary for reconstructing and sparsely encoding signals from measurements without phase. Specifically, we consider the task of estimating a two-dimensional image from squared-magnitude measurements of a complex-valued linear transformation of the original image. Several recent phase retrieval algorithms exploit underlying sparsity of the unknown signal in order to improve recovery performance. In this work, we consider such a sparse signal prior in the context of phase retrieval, when the sparsifying dictionary is not known in advance. Our algorithm jointly reconstructs the unknown

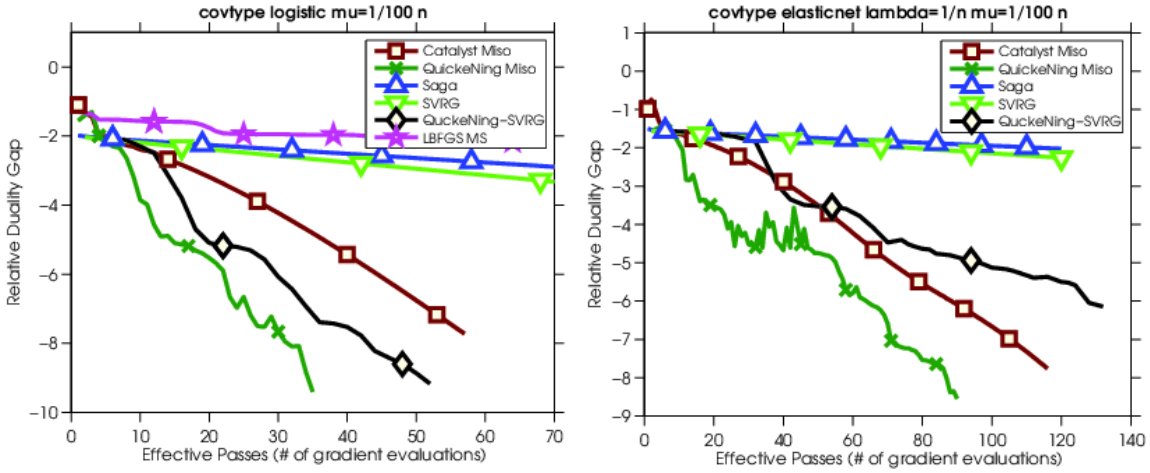

Figure 14. Relative duality gap for different number of passes performed over dataset covtype.

signal—possibly corrupted by noise—and learns a dictionary such that each patch of the estimated image can be sparsely represented. Numerical experiments demonstrate that our approach can obtain significantly better reconstructions for phase retrieval problems with noise than methods that cannot exploit such "hidden" sparsity. Moreover, on the theoretical side, we provide a convergence result for our method.

8. Bilateral Contracts and Grants with Industry

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

Participants: Cordelia Schmid, Karteek Alahari, Yang Hua.

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 [19] to videos and to develop an approach for action recognition based on temporal pose based on appropriate 3D features.

8.4. Google

Participants: Karteek Alahari, Cordelia Schmid.

We received a Google Faculty Research Award in 2015. The objective is to interpret video semantically in the presence of weak supervision. We will focus on answering questions such as *who* is in the scene, *what* they are doing, and *when* exactly did they perform their action(s). We propose to develop models for detection and recognition of objects and actions learned from minimally annotated training data.

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 semantic segmentation, and learning structure 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 will donate a state-of-the-art server with 8 GPUs.

8.6. MBDA

Participants: Jakob Verbeek, Julien Bardonnet.

Since 2004 we have collaborated with MBDA on a variety of subjects, namely object detection, tracking and matching. Several PhD students have been funded by MBDA, and code has been transferred which is integrated in products. Our collaboration resulted in 2010 in the award of the MBDA prize for innovation. Since May 2015 we have one engineer funded by MBDA working on incremental learning of object detection models. The goal is to take pre-existing vehicle models, and to quickly adapt them to new images of these vehicles when they are acquired in the field.

8.7. Xerox Research Center Europe

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

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.

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 Physionomie

Participants: Jakob Verbeek, Shreyas Saxena, Guosheng Hu.

Face recognition is nowadays an important technology in many applications ranging from tagging people in photo albums, to surveillance, and law enforcement. In this 3-year project (2013–2016) the goal is to broaden the scope of usefulness of face recognition to situations where high quality images are available in a dataset of known individuals, which have to be identified in relatively poor quality surveillance footage. To this end we will develop methods that can compare faces despite an asymmetry in the imaging conditions, as well as methods that can help searching for people based on facial attributes (old/young, male/female, etc.). The tools will be evaluated by law-enforcement professionals. The participants of this project are: Morpho, SensorIT, Université de Caen, Université de Strasbourg, Fondation pour la Recherche Stratégique, Préfecture de Police, Service des Technologies et des Systèmes d'Information de la Sécurité Intérieure, and Thoth. The project ended in June 2016.

9.2.2. 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].

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.

9.2.3. 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, Vicky Kalogeiton, Konstantin Shmelkov, Daan Wynen, Xiaojiang Peng.

The ERC advanced grant ALLEGRO started in April 2013 for a duration of five years. 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. EU Marie Curie project: Egovision4health

Participants: Grégory Rogez, Cordelia Schmid.

After the 2-year outgoing phase hosted by the University of California, Irvine, G. Rogez spent the return (and final) phase of the project in the team. In 2015, he analyzed functional object manipulations focusing on fine-grained hand-object interactions and created a large dataset of 12000 RGB-D images covering 71 everyday grasps in natural interactions. This Grasp UNderstanding dataset (GUN-71) has been made publicly available in 2016 (http://www.gregrogez.net/research/egovision4health/gun-71/). In the last period of the fellowship, G. Rogez and C. Schmid addressed the more general problem of full-body 3D pose estimation in third-person images. They developed a new data synthesis technique to generate large-scale (2 millions images) training data that were later used to train Deep Convolutional Neural Networks. The collaboration resulted in a publication [19]. Dataset, code and models will be released soon.

9.4. International Initiatives

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

9.4.1.1. GAYA: Semantic and Geometric Models for Video Interpretation

We have formed an associate team GAYA, with the primary goal of interpreting 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. Cordelia Schmid and Karteek Alahari are involved in this associate team.

9.4.2. Inria International Partners

9.4.2.1. Informal International Partners

- **University of Edinburgh:** C. Schmid collaborates with V. Ferrari, associate professor at university of Edinburgh. Vicky Kalogeiton started a co-supervised PhD in September 2013; she is bi-localized between Uni. Edinburgh and Inria. Her subject is the automatic learning of object representations in videos. The collaboration resulted in a joint publication in IEEE PAMI [7]
- **MPI Tübingen:** C. Schmid collaborates with M. Black, a research director at MPI, starting in 2013. She spent one month at MPI in May 2016. End of 2015 she was award a Humbolt research award funding a long-term research project with colleagues at MPI. In 2016 the project resulted in the development of a large-scale synthetic human action dataset.

- **Technion:** J. Mairal started a collaboration with Yonina Eldar (Technion) and Andreas Tillmann (Darmstadt university) to develop dictionary learning techniques for phase retrieval. Their collaboration resulted in a paper accepted to the ICASSP'16 conference [22] and a paper accepted to IEEE Transaction on signal processing [12].
- **UC Berkeley:** This collaboration between Bin Yu, Jack Gallant, Yuval Benjamini, Adam Bloniarz, Yuansi Chen (UC Berkeley), and Julien Mairal (Inria Thoth) aims to discover the functionalities of areas of the visual cortex. We have introduced an image representation for area V4, adapting tools from computer vision to neuroscience data. The collaboration started when Julien Mairal was a post-doctoral researcher at UC Berkeley and is still ongoing.

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.
- **France-Berkeley fund:** Julien Mairal was awarded in 2014 a grant from the France-Berkeley fund for a project with Pr. Bin Yu (statistics department, UC Berkeley) on "Invariant image representations and high dimensional sparse estimation for neurosciences". The award amounts to 10,000 USD, from November 2014 to April 2016. The funds are meant to support scientific and scholarly exchanges and collaboration between the two teams.

9.5. International Research Visitors

9.5.1. Visits to International Teams

9.5.1.1. Research Stays Abroad

- H. Lin visited Microsoft Research at New York from September to December 2016, as part of the MSR-Inria joint centre collaboration.
- G. Chéron visited Microsoft Research at Cambridge from April to July 2016, 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

- G. Rogez. Co-organizer of the CVPR workshop on Observing and Understanding Hands in Action (HANDS 2016).
- J. Verbeek. Organized Symposium on Computer Vision and Deep Learning on June 9th in Grenoble with around 80 attendants.

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

- C. Schmid: area chair for ECCV'16, ICCV'17.
- J. Mairal: area chair for CVPR 2016, ECCV 2016, ICLR 2016 and NIPS 2016.
- J. Verbeek: tutorial chair for ECCV'16.

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, AISTATS.

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. Verbeek: Associate editor for Image and Vision Computing Journal, since 2011.
- J. Verbeek: Associate editor for the International Journal on Computer Vision, since 2014.
- 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.

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

- C. Schmid. Invited speaker at Large-scale Computer Vision Workshop in conjunction with NIPS'16, December 2016.
- C. Schmid. Keynote speaker at IEEE International Conference on Image Processing, Phoenix, September 2016.
- C. Schmid. Invited speaker at Robust Features Workshop in conjunction with CVPR'16, June 2016.
- C. Schmid. Invited speaker at collège de France seminar (chair of Yann LeCun), Mars 2016.
- C. Schmid. Invited speaker at the LIG (laboratoire d'informatique de Grenoble) keynote talks, February 2016.
- C. Schmid. Seminar at Google, Montain View, July 2016.
- C. Schmid. Seminar at "journées scientifiques Inria", June 2016.
- C. Schmid. Seminar at Karlsruhe Technology Institute, June 2016.
- C. Schmid. Seminar at MPI, Tübingen, April 2016.
- C. Schmid. Seminar at INSA Lyon, April 2016.
- C. Schmid. Seminar at New York University, January 2016.
- J. Verbeek. Invited speaker at NVIDIA GPU Technology Conference, Amsterdam, The Netherlands. September 2016.
- J. Verbeek. Seminar GREYC, University of Caen, France, December 2016.
- J. Verbeek. Seminar PSI team, department of Electrical Engineering (ESAT), University of Leuven, Belgium, October 2016.
- J. Mairal. Invited talk at the Dagstuhl seminar "New Directions for Learning with Kernels and Gaussian Processes", December 2016.
- J. Mairal. Invited talk at workshop Phi-Tab, Telecom ParisTech, November 2016.
- J. Mairal. Invited talk at Journées GDR-Isis, Telecom ParisTech, September 2016.
- J. Mairal. Invited talk at Journées MAS, Grenoble, France, August 2016.
- J. Mairal. Invited talk at ICCOPT, Tokyo, Japan, August 2016.
- J. Mairal. Seminar at UC Berkeley, EECS department, USA, March 2016.

- J. Mairal. Seminar at UBC Vancouver, Canada, February 2016.
- J. Mairal. Invited talk at the MIA'16 workshop, Paris, France, January 2016.
- K. Alahari. Seminar at Carnegie Mellon University, USA, July 2016.
- K. Alahari. invited talk at Mysore Park workshop on vision, language and AI, India, December 2016.
- G. Rogez. Invited talk at Journées CNRS-GDR Isis, Telecom Paris, May 2016.
- G. Rogez. Invited speaker at CVPR Tutorial on First-person Visual Sensing: Theory, Models, and Application, Las Vegas, June 2016.
- G. Rogez. Seminar at LIRMM, Université de Montpellier, December 2016.
- H. Lin. Seminar at New York University, USA, April 2016.
- H. Lin. Seminar at Princeton University, USA, April 2016.
- H. Lin. Invited talk at ICCOPT, Tokyo, Japan, August 2016.

10.1.5. Scientific Expertise

- C. Schmid is member of the PAMI-TC awards committee, and the PAMI-TC executive commitee.
- K. Alahari: reviewer for National Sciences and Engineering Research Council of Canada (NSERC), Canada, Agence Nationale de la Recherche (ANR), and Icelandic Research Fund (IRF), Iceland.
- J. Mairal: reviewer for ANR.

10.1.6. Research Administration

- C. Schmid is member of the "comité d'orientations scientifiques". Inria Grenoble, 2016.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

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

Doctorat: J. Mairal, Lecturer at the summer school MAESTRA, Ohrid, Macedonia.

Master : C. Schmid, "Object recognition and computer vision", 9H eqTD, M2, ENS Cachan, France.

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

Master : J. Verbeek and J. Mairal, "Kernel Methods for Statistical Learning", 27H eqTD, M2, ENSIMAG, Grenoble.

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

Master: J. Mairal, "Introduction to sparse estimation", 6H eq-TD, M2, PSL-ITI, France.

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

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

Licence: P. Weinzaepfel, "Introduction à UNIX et à la programmation en langage C", 67.5H TD, L1, DLST Grenoble.

10.2.2. Supervision

PhD: P. Weinzaepfel, Motion in action : optical flow estimation and action localization in videos, supervision 50% C. Schmid and 50% Z. Harchaoui, September 2016.

PhD: Y. Hua, Towards robust visual object tracking : proposal selection and occlusion reasoning, supervision 50% C. Schmid and 50% K. Alahari, June 2016.

PhD: A. Mishra, Understanding Text in Scene Images, supervision 50% K. Alahari and 50% Prof. C. V. Jawahar, November 2016.

PhD: P. Bojanowski, Learning to annotate dynamic video scenes, supervision 20% with J. Ponce, I. Laptev and J. Sivic, June 2016.

PhD: S. Saxena, Learning representations for visual recognition, supervision 95% J. Verbeek and 5% C. Schmid, December 2016.

10.2.3. Juries

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

C. Schmid: Makarand Tapaswi, June 2016, rapporteur, these, KIT Karlsruhe.

C. Schmid: Natalia Neverova, avril 2016, president, these, INSA Lyon.

K. Alahari: Guillaume Seguin, 2016, examinateur, these, Ecole Normale Superieure, Paris, France.

G. Rogez. Marta Salas, 2016, rapporteur, these, Universidad de Zaragoza, Spain.

G. Rogez. Tu-Hoa Pham, December 2016, examinateur, these, Univ. Montpellier.

J. Verbeek. Amir Ghodrati, October 2016, rapporteur, these, Univ. Leuven.

J. Verbeek. Binod Bhattarai, December 2016, rapporteur, these, Univ. Caen.

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] Y. HUA. *Towards robust visual object tracking : proposal selection and occlusion reasoning*, Université Grenoble Alpes, June 2016, https://tel.archives-ouvertes.fr/tel-01394943

[2] J. VERBEEK. *Machine learning solutions to visual recognition problems*, Grenoble 1 UGA - Université Grenoble Alpes, June 2016, Habilitation à diriger des recherches, https://hal.inria.fr/tel-01343391

[3] P. WEINZAEPFEL. *Motion in action : optical flow estimation and action localization in videos*, Université Grenoble Alpes, September 2016, https://tel.archives-ouvertes.fr/tel-01407258

Articles in International Peer-Reviewed Journals

[4] R. G. CINBIS, J. VERBEEK, C. SCHMID. *Approximate Fisher Kernels of non-iid Image Models for Image Categorization*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", June 2016, vol. 38, n^{o} 6, pp. 1084-1098 [*DOI :* 10.1109/TPAMI.2015.2484342], https://hal.inria.fr/hal-01211201

[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://hal.inria.fr/hal-01123482

[6] M. DOUZE, J. REVAUD, J. VERBEEK, H. JÉGOU, C. SCHMID. *Circulant temporal encoding for video retrieval and temporal alignment*, in "International Journal of Computer Vision", 2016, vol. 119, n^{o} 3, pp. 291–306, https://hal.inria.fr/hal-01162603

[7] V. KALOGEITON, V. FERRARI, C. SCHMID. *Analysing domain shift factors between videos and images for object detection*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", April 2016 [*DOI :* 10.1109/TPAMI.2016.2551239], https://hal.inria.fr/hal-01281069

[8] A. MISHRA, K. ALAHARI, C. JAWAHAR. *Enhancing Energy Minimization Framework for Scene Text Recognition with Top-Down Cues*, in "Computer Vision and Image Understanding", April 2016, vol. 145, pp. 30-42 [*DOI :* 10.1016/J.CVIU.2016.01.002], https://hal.inria.fr/hal-01263322

[9] 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", August 2016 [*DOI :* 10.1007/S11263-016-0924-3], https://hal.inria.fr/hal-01277109

[10] J. REVAUD, P. WEINZAEPFEL, Z. HARCHAOUI, C. SCHMID. *DeepMatching: Hierarchical Deformable Dense Matching*, in "International Journal of Computer Vision", 2016 [*DOI :* 10.1007/S11263-016-0908-3], https://hal.inria.fr/hal-01148432

[11] G. SHARMA, F. JURIE, C. SCHMID. *Expanded Parts Model for Semantic Description of Humans in Still Images*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", January 2017, vol. 39, n^{o} 1, pp. 87-101 [*DOI :* 10.1109/TPAMI.2016.2537325], https://hal.inria.fr/hal-01199160

[12] A. TILLMANN, Y. ELDAR, J. MAIRAL. *DOLPHIn - Dictionary Learning for Phase Retrieval*, in "IEEE Transactions on Signal Processing", December 2016, vol. 64, n^{o} 24, pp. 6485-6500, author's preprint version [*DOI :* 10.1109/TSP.2016.2607180], https://hal.inria.fr/hal-01387428

[13] H. WANG, D. ONEATA, J. VERBEEK, C. SCHMID. *A robust and efficient video representation for action recognition*, in "International Journal of Computer Vision", 2016, vol. 119, n^{o} 3, pp. 219–238 [*DOI :* 10.1007/S11263-015-0846-5], https://hal.inria.fr/hal-01145834

Invited Conferences

[14] Z. HARCHAOUI, A. JUDITSKY, D. OSTROVSKI. *Filtrage adaptatif par optimisation convexe*, in "Journées SMAI-MODE 2016", Toulouse, France, March 2016, https://hal.archives-ouvertes.fr/hal-01336268

International Conferences with Proceedings

[15] B. HAM, M. CHO, C. SCHMID, J. PONCE. *Proposal Flow*, in "CVPR 2016 - IEEE Conference on Computer Vision & Pattern Recognition", LAS VEGAS, United States, June 2016, https://hal.archives-ouvertes.fr/hal-01240281

[16] J. MAIRAL. *End-to-End Kernel Learning with Supervised Convolutional Kernel Networks*, in "Advances in Neural Information Processing Systems (NIPS)", Barcelona, France, December 2016, https://hal.inria.fr/hal-01387399

[17] A. MENSCH, J. MAIRAL, B. THIRION, G. VAROQUAUX. *Dictionary Learning for Massive Matrix Factorization*, in "International Conference on Machine Learning", New York, United States, Proceedings of the 33rd Internation Conference on Machine Learning, June 2016, vol. 48, pp. 1737–1746, https://hal.archives-ouvertes.fr/hal-01308934

[18] X. PENG, C. SCHMID. *Multi-region two-stream R-CNN for action detection*, in "ECCV 2016 - European Conference on Computer Vision", Amsterdam, Netherlands, Lecture Notes in Computer Science, Springer, October 2016, vol. 9908, pp. 744-759 [*DOI :* 10.1007/978-3-319-46493-0_45], https://hal.inria.fr/hal-01349107

[19] G. ROGEZ, C. SCHMID. *MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild*, in "Advances in Neural Information Processing Systems (NIPS)", Barcelona, Spain, December 2016, https://hal.inria.fr/hal-01389486

[20] S. SAXENA, J. VERBEEK. *Convolutional Neural Fabrics*, in "Advances in Neural Information Processing Systems (NIPS)", Barcelona, Spain, December 2016, https://hal.inria.fr/hal-01359150

[21] S. SAXENA, J. VERBEEK. *Heterogeneous Face Recognition with CNNs*, in "ECCV TASK-CV 2016 Workshops", Amsterdam, Netherlands, October 2016, https://hal.inria.fr/hal-01367455

[22] A. TILLMANN, Y. ELDAR, J. MAIRAL. *Dictionary learning from phaseless measurements*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)", Shanghai, China, IEEE, March 2016, pp. 4702-4706 [*DOI :* 10.1109/ICASSP.2016.7472569], https://hal.inria.fr/hal-01387416

[23] P. TOKMAKOV, K. ALAHARI, C. SCHMID. *Weakly-Supervised Semantic Segmentation using Motion Cues*, in "ECCV 2016 - European Conference on Computer Vision", Amsterdam, Netherlands, October 2016, https://hal.archives-ouvertes.fr/hal-01351135

Conferences without Proceedings

[24] P. LUC, C. COUPRIE, S. CHINTALA, J. VERBEEK. *Semantic Segmentation using Adversarial Networks*, in "NIPS Workshop on Adversarial Training", Barcelona, Spain, December 2016, https://hal.inria.fr/hal-01398049

[25] A. MENSCH, J. MAIRAL, G. VAROQUAUX, B. THIRION. *Subsampled online matrix factorization with convergence guarantees*, in "NIPS Workshop on Optimization for Machine Learning", Barcelone, Spain, December 2016, https://hal.archives-ouvertes.fr/hal-01405058

Other Publications

[26] A. BIETTI, J. MAIRAL. *Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite-Sum Structure*, October 2016, working paper; a short version has been accepted to the NIPS OPT2016 workshop, https://hal.inria.fr/hal-01375816

[27] G. HU, X. PENG, Y. YANG, T. HOSPEDALES, J. VERBEEK. *Frankenstein: Learning Deep Face Representations using Small Data*, April 2016, working paper or preprint, https://hal.inria.fr/hal-01306168

[28] H. LIN, J. MAIRAL, Z. HARCHAOUI. *QuickeNing: A Generic Quasi-Newton Algorithm for Faster Gradient-Based Optimization*, October 2016, working paper; a short version has been accepted to the NIPS workshop on optimization for machine learning 2016, https://hal.inria.fr/hal-01376079

[29] M. PEDERSOLI, T. LUCAS, C. SCHMID, J. VERBEEK. *Areas of Attention for Image Captioning*, November 2016, working paper or preprint, https://hal.inria.fr/hal-01428963

[30] P. TOKMAKOV, K. ALAHARI, C. SCHMID. *Learning Semantic Segmentation with Weakly-Annotated Videos*, July 2016, working paper or preprint, https://hal.inria.fr/hal-01292794

[31] P. TOKMAKOV, K. ALAHARI, C. SCHMID. *Learning Motion Patterns in Videos*, January 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01427480

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

[33] P. WEINZAEPFEL, X. MARTIN, C. SCHMID. *Towards Weakly-Supervised Action Localization*, May 2016, working paper or preprint, https://hal.inria.fr/hal-01317558