IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université Joseph Fourier (Grenoble)

Activity Report 2014

Project-Team LEAR

Learning and recognition in vision

IN COLLABORATION WITH: Laboratoire Jean Kuntzmann (LJK)

RESEARCH CENTER
Grenoble - Rhône-Alpes

THEME
Vision, perception and multimedia interpretation

Table of contents

Project-Team LEAR

Keywords: Computer Vision, Machine Learning, Video, Recognition

Creation of the Project-Team: 2003 July 01.

1. Members

Research Scientists

Cordelia Schmid [Team leader, Inria, Senior Researcher, HdR]
Karteek Alahari [Inria, Inria Starting research position, from Sep 2013 until Aug 2016]
Zaid Harchaoui [Inria, Researcher]
Julien Mairal [Inria, Researcher, "en détachement du Corps des Mines"]
Jakob Verbeek [Inria, Researcher]

Engineers

Matthijs Douze [Inria, SED 40 %]
Clement Leray [Inria, funded by FP7 AXES project, from Sep 2013 until Sep 2014]
Xavier Martin [Inria, funded by FP7 AXES project, from Oct 2014 until Oct 2015]

PhD Students

Guilhem Cheron [Ens, funded by MSR/Inria, from Oct 2014 until Oct 2017, co-supervision with I. Laptev]
Nicolas Chesneau [Univ. Grenoble, funded by European Research Council, ALLEGRO project, from Jul 2014 until Sep 2017]
Ramazan Cinbis [Univ. Grenoble, funded by FP7 AXES project, from Oct 2010 until March 2014]
Yang Hua [Univ. Grenoble, funded by MSR/Inria joint lab, from Jan 2013 until Dec 2015]
Vicky Kalogeiton [Univ. Edinburgh, European Research Council, co-supervision with V. Ferrari, from Sep 2013 until Dec 2016]
Hongzhou Lin [Univ. Grenoble, funded by Université Joseph Fourier, from Apr 2014 until Sep 2017]
Dan Oneata [Univ. Grenoble, funded by FP7 AXES project, from Oct 2011 until Jul 2015]
Mattis Paulin [Univ. Grenoble, funded by DGA, from Apr 2013 until Apr 2016]
Federico Pierucci [Univ. Grenoble, funded by Université Joseph Fourier, from Jan 2012 until Sep 2015]
Danila Potapov [Univ. Grenoble, FP7 AXES project and Quaero, from Sep 2011 until Mar 2015]
Shreyas Saxena [Univ. Grenoble, ANR PHYSIONOMIE project, from Feb 2013 until Sep 2016]
Pavel Tokmakov [Univ. Grenoble, funded by European Research Council, ALLEGRO project, from Sep 2014 until Sep 2017]
Philippe Weinzaepfel [Univ. Grenoble, funded by Université Joseph Fourier, from Nov 2012 until Sep 2015]

Post-Doctoral Fellows

Anoop Cherian [Inria, funded by FP7 AXES project, from Nov 2012 until Feb 2015]
Albert Gordo [Inria, funded by MBDA, from Aug 2012 until Jan 2014]
Piotr Koniusz [Inria, Inria Fellowship, from Jul 2013 until Mar 2015]
Yuri Maximov [Univ. Grenoble, from Jan 2014 until Dec 2014]
Jerome Revaud [Inria, funded by European Research Council, ALLEGRO project, from Jun 2011 until Dec 2015]
Heng Wang [Inria, funded by FP7 AXES project, from July 2012 until April 2014]

Visiting Scientist

Thanh Tam Le [Kyoto University, Feb 2014]

Administrative Assistant

Nathalie Gillot [Inria]

Others

Lucas Claude [Inria, intern, from Apr 2014 until Jul 2014]

Quentin Cormier [Inria, intern, from Jun 2014 until Jul 2014]
Dmitry Ostrovsky [Univ. Grenoble, intern, from Feb 2014 until Jul 2014]
Albin Toulisse [Inria, intern, from Jun 2014 until Jul 2014]
Daan Wynen [Inria, intern, from Nov 2014]

2. Overall Objectives

2.1. Introduction

LEAR's main focus is learning-based approaches to visual object recognition and scene interpretation. Understanding the content of everyday images and videos is one of the fundamental challenges of computer vision, and our approach is based on developing state-of-the-art visual models along with machine learning and statistical modeling techniques.

Key problems in computer vision are robust image and video representations. We have over the past years developed robust image descriptions invariant to different image transformations and illumination changes. We have more recently concentrated on the problem of robust object and videos representations. The descriptions can be either low-level or build on mid or high-level descriptions.

In order to deal with large quantities of visual data and to extract relevant information automatically, we develop machine learning techniques that can handle the huge volumes of data that image and video collections contain. We also want to handle noisy training data and to combine vision with textual data as well as to capture enough domain information to allow generalization from just a few images rather than having to build large, carefully marked-up training databases. Furthermore, the selection and coupling of image descriptors and learning techniques is today often done by hand, and one significant challenge is the automation of this process, for example using automatic feature learning.

LEAR's main research areas are:

- **Large-scale image search and categorization.** Searching and categorizing large collections of images and videos becomes more and more important as the amount of digital information available explodes. The two main issues to be solved are (1) the development of efficient algorithms for very large image collections and (2) the definition of semantic relevance. Visual recognition is currently reaching a point where models for thousands of object classes are learned. To further improve the performance, we will need to work on new learning techniques that take into account the different misclassification costs, e.g., classifying a bus as a car is clearly better than classifying it as a horse. A solution to these problems will be applicable to many different real-world problems, as for example image-based internet search.
- **Statistical modeling and machine learning for visual recognition.** Our work on statistical modeling and machine learning is aimed mainly at developing techniques to improve visual recognition. This includes both the selection, evaluation and adaptation of existing methods, and the development of new ones designed to take vision specific constraints into account. Particular challenges include: (i) the need to deal with the huge volumes of data that image and video collections contain; (ii) the need to handle "noisy" training data, i.e., to combine vision with textual data; and (iii) the need to capture enough domain information to allow generalization from just a few images rather than having to build large, carefully marked-up training databases.
- **Recognizing humans and their actions.** Humans and their activities are one of the most frequent and interesting subjects in images and videos, but also one of the hardest to analyze owing to the complexity of the human form, clothing and movements. Our research aims at developing robust descriptors to characterize humans and their movements. This includes methods for identifying humans as well as their pose in still images as well as videos. Furthermore, we investigate appropriate descriptors for capturing the temporal motion information characteristic for human actions. Video, furthermore, permits to easily acquire large quantities of data often associated with text obtained from transcripts. Methods will use this data to automatically learn actions despite the noisy labels.

- **Automatic learning of visual models.** Our goal is to advance the state of visual modeling given weakly labeled images and videos. We will depart from the essentially rigid (or piecewise-rigid) object models typically used in object recognition and detection tasks by introducing flexible models assembled from local image evidence. We will use the abundant data to leverage the underlying latent structure between features, classes and examples and to build efficient algorithms to iteratively train multilayer architectures that adapt to an increasing pool of labeled examples. This will allow us to capture the evolving appearance of objects under changes in viewpoint, combine detection and tracking using motion information and, perhaps more importantly, learn the dynamic relationship between object categories, people, and scene context.

3. Research Program

3.1. Image features and descriptors and robust correspondence

Reliable image features are a crucial component of any visual recognition system. Despite much progress, research is still needed in this area. Elementary features and descriptors suffice for a few applications, but their lack of robustness and invariance puts a heavy burden on the learning method and the training data, ultimately limiting the performance that can be achieved. More sophisticated descriptors allow better inter-class separation and hence simpler learning methods, potentially enabling generalization from just a few examples and avoiding the need for large, carefully engineered training databases.

The feature and descriptor families that we advocate typically share several basic properties:

- **Locality and redundancy:** For resistance to variable intra-class geometry, occlusions, changes of viewpoint and background, and individual feature extraction failures, descriptors should have relatively small spatial support and there should be many of them in each image. Schemes based on collections of image patches or fragments are more robust and better adapted to object-level queries than global whole-image descriptors. A typical scheme thus selects an appropriate set of image fragments, calculates robust appearance descriptors over each of these, and uses the resulting collection of descriptors as a characterization of the image or object (a "bag-of-features" approach – see below).
- **Photometric and geometric invariance:** Features and descriptors must be sufficiently invariant to changes of illumination and image quantization and to variations of local image geometry induced by changes of viewpoint, viewing distance, image sampling and by local intra-class variability. In practice, for local features geometric invariance is usually approximated by invariance to Euclidean, similarity or affine transforms of the local image.
- **Repeatability and salience:** Fragments are not very useful unless they can be extracted reliably and found again in other images. Rather than using dense sets of fragments, we often focus on local descriptors based at particularly salient points – "keypoints" or "points of interest". This gives a sparser and thus potentially more efficient representation, and one that can be constructed automatically in a preprocessing step. To be useful, such points must be accurately relocalizable in other images, with respect to both position and scale.
- **Informativeness:** Notwithstanding the above forms of robustness, descriptors must also be informative in the sense that they are rich sources of information about image content that can easily be exploited in scene characterization and object recognition tasks. Images contain a lot of variety so high-dimensional descriptions are required. The useful information should also be manifest, not hidden in fine details or obscure high-order correlations. In particular, image formation is essentially a spatial process, so relative position information needs to be made explicit, e.g. using local feature or context style descriptors.

Partly owing to our own investigations, features and descriptors with some or all of these properties have become popular choices for visual correspondence and recognition, particularly when large changes of viewpoint may occur. One notable success to which we contributed is the rise of "bag-of-features" methods for visual object recognition. These characterize images by their (suitably quantized or parametrized) global distributions of local descriptors in descriptor space. The representation evolved from texton based methods in texture analysis. Despite the fact that it does not (explicitly) encode much spatial structure, it turns out to be surprisingly powerful for recognizing more structural object categories.

Our current research on local features is focused on creating detectors and descriptors that are better adapted to describe object classes, on incorporating spatial neighborhood and region constraints to improve informativeness relative to the bag-of-features approach, and on extending the scheme to cover different kinds of locality. Current research also includes the development and evaluation of local descriptors for video, and associated detectors for spatio-temporal content.

3.2. Statistical modeling and machine learning for image analysis

We are interested in learning and statistics mainly as technologies for attacking difficult vision problems, so we take an eclectic approach, using a broad spectrum of techniques ranging from classical statistical generative and discriminative models to modern kernel, margin and boosting based approaches. Hereafter we enumerate a set of approaches that address some problems encountered in this context.

- Parameter-rich models and limited training data are the norm in vision, so overfitting needs to be estimated by cross-validation, information criteria or capacity bounds and controlled by regularization, model and feature selection.
- Visual descriptors tend to be high-dimensional and redundant, so we often preprocess data to reduce it to more manageable terms using dimensionality reduction techniques including PCA and its non-linear variants, latent structure methods such as Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA), and manifold methods such as Isomap/LLE.
- To capture the shapes of complex probability distributions over high-dimensional descriptor spaces, we either fit mixture models and similar structured semi-parametric probability models, or reduce them to histograms using vector quantization techniques such as K-means or latent semantic structure models.
- Missing data is common owing to unknown class labels, feature detection failures, occlusions and intra-class variability, so we need to use data completion techniques based on variational methods, belief propagation or MCMC sampling.
- Weakly labeled data is also common – for example one may be told that a training image contains an object of some class, but not where the object is in the image – and variants of unsupervised, semi-supervised and co-training are useful for handling this. In general, it is expensive and tedious to label large numbers of training images so less supervised data mining style methods are an area that needs to be developed.
- On the discriminative side, machine learning techniques such as Support Vector Machines, Relevance Vector Machines, and Boosting, are used to produce flexible classifiers and regression methods based on visual descriptors.
- Visual categories have a rich nested structure, so techniques that handle large numbers of classes and nested classes are especially interesting to us.
- Images and videos contain huge amounts of data, so we need to use algorithms suited to large-scale learning problems.

3.3. Visual recognition and content analysis

Current progress in visual recognition shows that combining advanced image descriptors with modern learning and statistical modeling techniques is producing significant advances. We believe that, taken together and tightly integrated, these techniques have the potential to make visual recognition a mainstream technology that is regularly used in applications ranging from visual navigation through image and video databases to human-computer interfaces and smart rooms.

The recognition strategies that we advocate make full use of the robustness of our invariant image features and the richness of the corresponding descriptors to provide a vocabulary of base features that already goes a long way towards characterizing the category being recognized. Trying to learn everything from scratch using simpler, non-invariant features would require far too much data: good learning cannot easily make up for bad features. The final classifier is thus responsible "only" for extending the base results to larger amounts of intra-class and viewpoint variation and for capturing higher-order correlations that are needed to fine tune the performance.

That said, learning is not restricted to the classifier and feature sets can not be designed in isolation. We advocate an end-to-end engineering approach in which each stage of the processing chain combines learning with well-informed design and exploitation of statistical and structural domain models. Each stage is thoroughly tested to quantify and optimize its performance, thus generating or selecting robust and informative features, descriptors and comparison metrics, squeezing out redundancy and bringing out informativeness.

4. Application Domains

4.1. Application Domains

A solution to the general problem of visual recognition and scene understanding will enable a wide variety of applications in areas including human-computer interaction, retrieval and data mining, medical and scientific image analysis, manufacturing, transportation, personal and industrial robotics, and surveillance and security. With the ever expanding array of image and video sources, visual recognition technology is likely to become an integral part of many information systems. A complete solution to the recognition problem is unlikely in the near future, but partial solutions in these areas enable many applications. LEAR's research focuses on developing basic methods and general purpose solutions rather than on a specific application area. Nevertheless, we have applied our methods in several different contexts.

Semantic-level image and video access. This is an area with considerable potential for future expansion owing to the huge amount of visual data that is archived. Besides the many commercial image and video archives, it has been estimated that as much as 96% of the new data generated by humanity is in the form of personal videos and images [1], and there are also applications centering on on-line treatment of images from camera equipped mobile devices (e.g. navigation aids, recognizing and answering queries about a product seen in a store). Technologies such as MPEG-7 provide a framework for this, but they will not become generally useful until the required mark-up can be supplied automatically. The base technology that needs to be developed is efficient, reliable recognition and hyperlinking of semantic-level domain categories (people, particular individuals, scene type, generic classes such as vehicles or types of animals, actions such as football goals, etc).

Visual (example based) search. The essential requirement here is robust correspondence between observed images and reference ones, despite large differences in viewpoint or malicious attacks of the images. The reference database is typically large, requiring efficient indexing of visual appearance. Visual search is a key component of many applications. One application is navigation through image and video datasets, which is essential due to the growing number of digital capture devices used by industry and individuals. Another application that currently receives significant attention is copyright protection. Indeed, many images and videos covered by copyright are illegally copied on the Internet, in particular on peer-to-peer networks or on the so-called user-generated content sites such as Flickr, YouTube or DailyMotion. Another type of application is the detection of specific content from images and videos, which can, for example, be used for finding product related information given an image of the product.

[1] http://www.sims.berkeley.edu/research/projects/how-much-info/summary.html

Automated object detection. Many applications require the reliable detection and localization of one or a few object classes. Examples are pedestrian detection for automatic vehicle control, airplane detection for military applications and car detection for traffic control. Object detection has often to be performed in less common imaging modalities such as infrared and under significant processing constraints. The main challenges are the relatively poor image resolution, the small size of the object regions and the changeable appearance of the objects.

5. New Software and Platforms

5.1. Yael library

Participants: Matthijs Douze [correspondant], Herve Jegou [TEXMEX Team Inria Rennes].

Yael [14] is a library with Matlab and Python bindings providing optimized (multi-threaded, Blas/Lapack, low level optimization) implementations of functions useful in vision and machine learning such as k-means, GMM, exact nearest neighbor search and Fisher vector computation.

In 2014, it was extended to include a generic inverted file implementation, that can accomodate any type of signature that refines the similarity computation between documents. The Fisher vector computation code was also optimized.

5.2. SPArse Modeling Software (SPAMS)

Participants: Julien Mairal [correspondant], Yuansi Chen, Zaid Harchaoui.

SPAMS v2.5 was released as open-source software in May 2014 (v1.0 was released in September 2009). It is an optimization toolbox implementing algorithms to address various machine learning and signal processing problems involving

- Dictionary learning and matrix factorization (NMF, sparse PCA, ...);
- Solving medium-scale sparse decomposition problems with LARS, coordinate descent, OMP, SOMP, proximal methods;
- Solving large-scale sparse estimation problems with stochastic optimization;
- Solving structured sparse decomposition problems (sparse group lasso, tree-structured regularization, structured sparsity with overlapping groups,...).

The software and its documentation are available at http://spams-devel.gforge.inria.fr/.

This year, we added new functionalities to the toolbox. The implementation of archetypal analysis corresponding to the paper [9] was added.

5.3. FlipFlop: Fast Lasso-based Isoform Prediction as a Flow Problem

Participants: Elsa Bernard [Institut Curie, Ecoles des Mines-ParisTech], Laurent Jacob [CNRS, LBBE Laboratory], Julien Mairal [correspondant], Jean-Philippe Vert [Institut Curie, Ecoles des Mines-ParisTech].

FlipFlop is an open-source software, implementing a fast method for de novo transcript discovery and abundance estimation from RNA-Seq data [4]. It differs from classical approaches such as Cufflinks by simultaneously performing the identification and quantitation tasks using a penalized maximum likelihood approach, which leads to improved precision/recall. Other software taking this approach have an exponential complexity in the number of exons of a gene. We use a novel algorithm based on network flow formalism, which gives us a polynomial runtime. In practice, FlipFlop was shown to outperform penalized maximum likelihood based softwares in terms of speed and to perform transcript discovery in less than 1/2 second for large genes.

FlipFlop 1.4.1 is a user friendly bioconductor R package, which was released in October 2014. It is freely available on the Bioconductor website under a GPL licence: http://bioconductor.org/packages/release/bioc/html/flipflop.html.

5.4. DeepFlow

Participants: Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, Cordelia Schmid.

We developed a package for the "deep flow" algorithm. "Deep flow" combines a standard variational framework with a our new matching algorithm "deep matching". The code for "deep matching" is in python and the code for "deep flow" in C. Both of them are available on-line at http://lear.inrialpes.fr/src/deepmatching. Note that the run time is a few seconds per images pair, which is less than for most other methods. The latest release was published in March 2014.

5.5. Mixing Body-Part Sequences for Human Pose Estimation

Participants: Cherian Anoop, Mairal Julien, Alahari Karteek, Schmid Cordelia.

The code corresponding to the publication [11] has been released as an open-source MATLAB package along with a dataset for human pose estimation in videos called "Poses in the Wild". It is available at http://lear.inrialpes.fr/research/posesinthewild/#dataset. This dataset has 30 video sequences generated from three Hollywood movies, namely "Forrest Gump", "The Terminal", and "Cast Away". Each sequence has approximately 30 frames and is manually annotated for human upper-body keypoints, namely (i) neck, (ii) left and right shoulders, (iii) left and right elbows, (iv) left and right wrists, and (v) mid-torso. In comparison to earlier evaluation datasets publicly available for this problem, Poses in the Wild is significantly more representative of real-world scenarios with background clutter, body-part occlusions, and severe camera motion.

5.6. Image Transformation Pursuit

Participants: Mattis Paulin, Jerome Revaud, Zaid Harchaoui, Florent Perronnin [XRCE], Cordelia Schmid.

This is an open-source software package corresponding to the papers [19], [23], available here http://lear.inrialpes.fr/people/paulin/projects/ITP/. The code has three main purposes. Starting from input images, it can be used to generate transformed versions to use as "virtual examples". It implements the main algorithm of the article (ITP), performing an automatic selection of a small set of transformations in order to improve classification performance. Lastly, it provides a complete classification framework, allowing to train and test a classifier on an image dataset.

5.7. Convolutional Kernel Networks

Participants: Julien Mairal, Piotr Koniusz, Zaid Harchaoui, Cordelia Schmid.

This is an open-source software package corresponding to the paper [16], available at http://ckn.gforge.inria.fr/. In this software package, convolutional neural networks are learned in an unsupervised manner. We control what the non-linearities of the network are really doing: the network tries to approximate the kernel map of a reproducing kernel.

5.8. EpicFlow

Participants: Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, Cordelia Schmid.

We developed a package for the EpicFlow method [29]. EpicFlow computes a dense correspondence field by performing a sparse-to-dense interpolation from an initial sparse set of matches, leveraging contour cues using an edge-aware geodesic distance. The resulting dense correspondence field is fed as an initial optical flow estimate to a one-level variational energy minimization. The code is written in C/C++ and is available at http://lear.inrialpes.fr/src/epicflow.

6. New Results

6.1. Highlights of the Year

- Cordelia Schmid received the **Longuet-Higgins prize** for fundamental contributions in computer vision that have withstood the test of time, 2014.
- We participated to the **Trecvid 2014 Multimedia Event Detection** challenge. We ranked first on one of the four tracks (Ad-hoc training videos with 10 examples per class).
- We participated to the **THUMOS 2014 challenge**. We obtained top ranked results in the localization track of the Thumos 2014 Action Recognition Challenge. The goal of the challenge is to evaluate large-scale action recognition in natural settings.

6.2. Visual recognition in images

6.2.1. Multi-fold MIL Training for Weakly Supervised Object Localization

Participants: Ramazan Cinbis, Cordelia Schmid, Jakob Verbeek.

Object category localization is a challenging problem in computer vision. Standard supervised training requires bounding box annotations of object instances. This time-consuming annotation process is sidestepped in weakly supervised learning. In this case, the supervised information is restricted to binary labels that indicate the absence/presence of object instances in the image, without their locations. In [13], we follow a multiple-instance learning approach that iteratively trains the detector and infers the object locations in the positive training images. Our main contribution is a multi-fold multiple instance learning procedure, which prevents training from prematurely locking onto erroneous object locations. This procedure is particularly important when high-dimensional representations, such as the Fisher vectors, are used. We present a detailed experimental evaluation using the PASCAL VOC 2007 and 2010 datasets. Compared to state-of-the-art weakly supervised detectors, our approach better localizes objects in the training images, which translates into improved detection performance. Figure 1 illustrates the iterative object localization process on several example images.

A journal paper is currently in preparation in which extends [13] by adding experiments with CNN features, and a refinement procedure for the object location inference. These additions improve over related work that has appeared since the publication of the original paper.

6.2.2. Transformation Pursuit for Image Classification

Participants: Mattis Paulin, Jerome Revaud, Zaid Harchaoui, Florent Perronnin [XRCE], Cordelia Schmid.

In this work [19], [23], we use data augmentation (see Fig 2 for examples) to improve image classification performances in a large-scale context. A simple approach to learning invariances in image classification consists in augmenting the training set with transformed versions of the original images. However, given a large set of possible transformations, selecting a compact subset is challenging. Indeed, all transformations are not equally informative and adding uninformative transformations increases training time with no gain in accuracy. We propose a principled algorithm – Image Transformation Pursuit (ITP) – for the automatic selection of a compact set of transformations. ITP works in a greedy fashion, by selecting at each iteration the one that yields the highest accuracy gain. ITP also allows to efficiently explore complex transformations, that combine basic transformations. We report results on two public benchmarks: the CUB dataset of bird images and the ImageNet 2010 challenge. Using Fisher Vector representations, we achieve an improvement from 28.2% to 45.2% in top-1 accuracy on CUB, and an improvement from 70.1% to 74.9% in top-5 accuracy on ImageNet. We also show significant improvements for deep convnet features: from 47.3% to 55.4% on CUB and from 77.9% to 81.4% on ImageNet.

6.2.3. Convolutional Kernel Networks

Participants: Julien Mairal, Piotr Koniusz, Zaid Harchaoui, Cordelia Schmid.

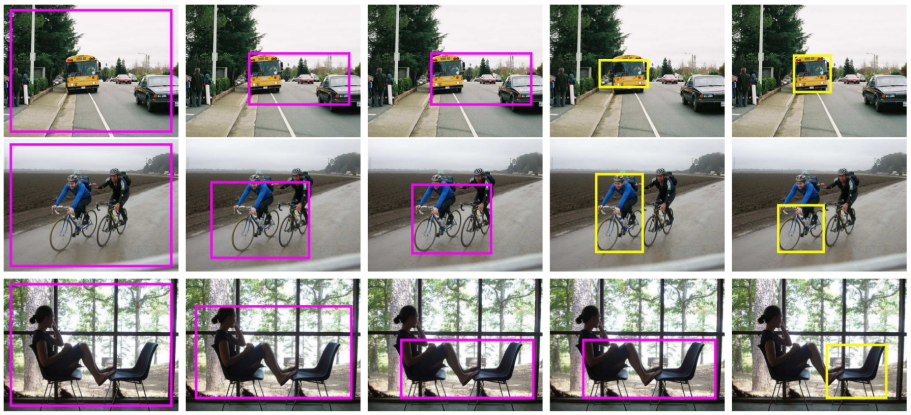


Figure 1. Illustrattion of our iterative object localization process on several example images, from initialization (left) to final localization (right). Yellow bounding boxes indicate that the object location hypothesis is in agreement with the ground-truth, for pink boxes the hypothesis is incorrect.

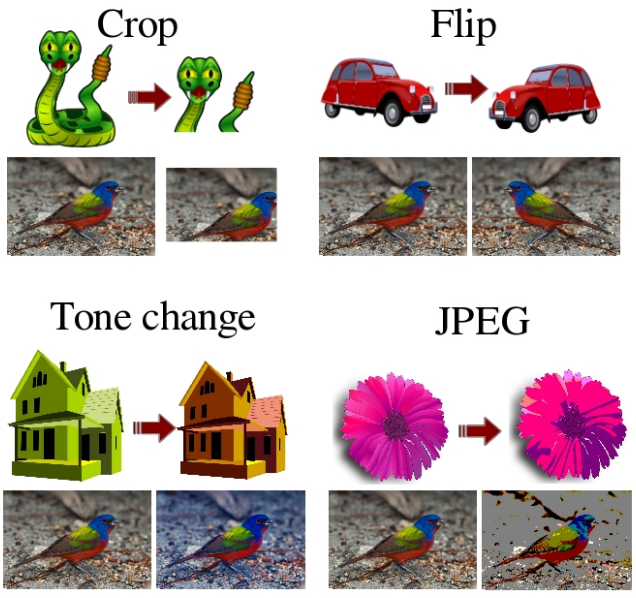
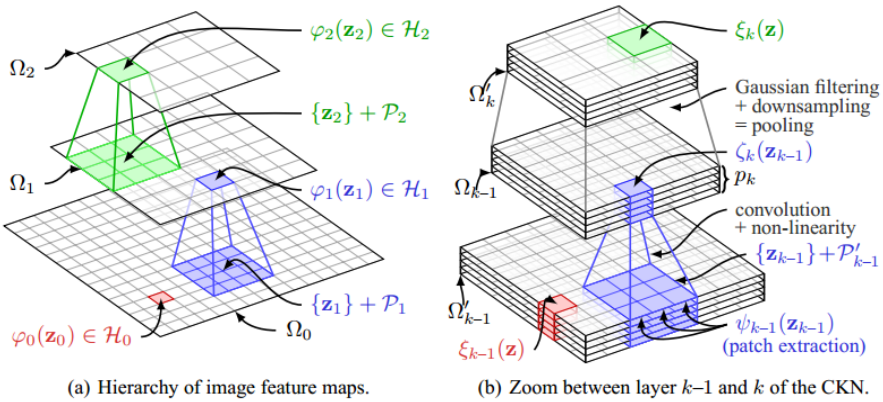


Figure 2. Examples of transformations used in [19], [23].

An important goal in visual recognition is to devise image representations that are invariant to particular transformations. In this paper [16] we address this goal with a new type of convolutional neural network (CNN) whose invariance is encoded by a reproducing kernel. Unlike traditional approaches where neural networks are learned either to represent data or for solving a classification task, our network learns to approximate the kernel feature map on training data. Such an approach enjoys several benefits over classical ones. First, by teaching CNNs to be invariant, we obtain simple network architectures that achieve a similar accuracy to more complex ones, while being easy to train and robust to overfitting. Second, we bridge a gap between the neural network literature and kernels, which are natural tools to model invariance. We evaluate our methodology on visual recognition tasks where CNNs have proven to perform well, e.g., digit recognition with the MNIST dataset, and the more challenging CIFAR-10 and STL-10 datasets, where our accuracy is competitive with the state of the art. Figure 3 illustrates the architecture of our network.



(a) Hierarchy of image feature maps. (b) Zoom between layer k–1 and k of the CKN.

Figure 3. Left: concrete representation of the successive layers for the multilayer convolutional kernel. Right: one layer of the convolutional neural network that approximates the kernel.

6.2.4. Scene Text Recognition and Retrieval for Large Lexicons

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

In [21], we propose a framework for recognition and retrieval tasks in the context of scene text images. In contrast to many of the recent works, we focus on the case where an image-specific list of words, known as the small lexicon setting, is unavailable. We present a conditional random field model defined on potential character locations and the interactions between them. Observing that the interaction potentials computed in the large lexicon setting are less effective than in the case of a small lexicon, we propose an iterative method, which alternates between finding the most likely solution and refining the interaction potentials. We evaluate our method on public datasets and show that it improves over baseline and state-of-the-art approaches. For example, we obtain nearly 15% improvement in recognition accuracy and precision for our retrieval task over baseline methods on the IIIT-5K word dataset, with a large lexicon containing 0.5 million words.

6.2.5. On Learning to Localize Objects with Minimal Supervision

Participants: Hyun On Song [UC Berkeley], Ross Girschick [UC Berkeley], Stefanie Jegelka [UC Berkeley], Julien Mairal, Zaid Harchaoui, Trevor Darrell [UC Berkeley].

Learning to localize objects with minimal supervision is an important problem in computer vision, since large fully annotated datasets are extremely costly to obtain. In this paper [22], we propose a new method that achieves this goal with only image-level labels of whether the objects are present or not. Our approach combines a discriminative submodular cover problem for automatically discovering a set of positive object windows with a smoothed latent SVM formulation. The latter allows us to leverage efficient quasiNewton optimization techniques. Experimental results are presented in Figure 4.

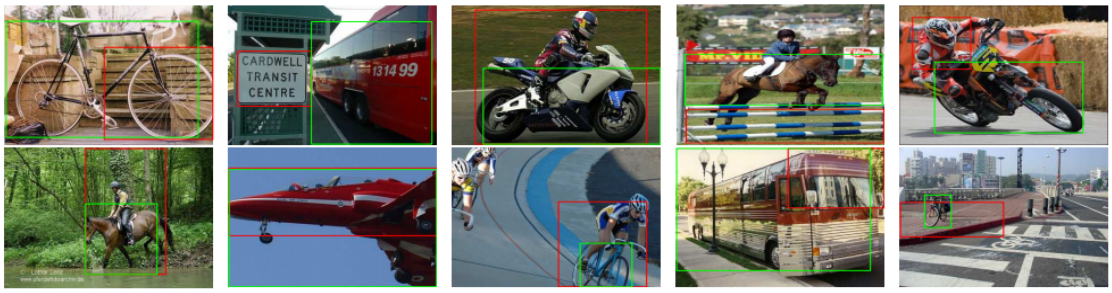


Figure 4. Visualization of some common failure cases of constructed positive windows by (Siva et al., 2012) vs our method. Red bounding boxes are constructed positive windows from (Siva et al., 2012). Green bounding boxes are constructed positive windows from our method.

6.2.6. Good Practice in Large-Scale Learning for Image Classification

Participants: Zeynep Akata, Florent Perronnin [XRCE], Zaid Harchaoui, Cordelia Schmid.

In this paper [3], we benchmark several SVM objective functions for large-scale image classification. We consider one-vs-rest, multi-class, ranking, and weighted approximate ranking SVMs. A comparison of online and batch methods for optimizing the objectives shows that online methods perform as well as batch methods in terms of classification accuracy, but with a significant gain in training speed. Using stochastic gradient descent, we can scale the training to millions of images and thousands of classes. Our experimental evaluation shows that ranking-based algorithms do not outperform the one-vs-rest strategy when a large number of training examples are used. Furthermore, the gap in accuracy between the different algorithms shrinks as the dimension of the features increases. We also show that learning through cross-validation the optimal rebalancing of positive and negative examples can result in a significant improvement for the one-vs-rest strategy. Finally, early stopping can be used as an effective regularization strategy when training with online algorithms. Following these "good practices", we were able to improve the state-of-the-art on a large subset of 10K classes and 9M images of ImageNet from 16.7% Top-1 accuracy to 19.1%.

6.3. Learning and statistical models

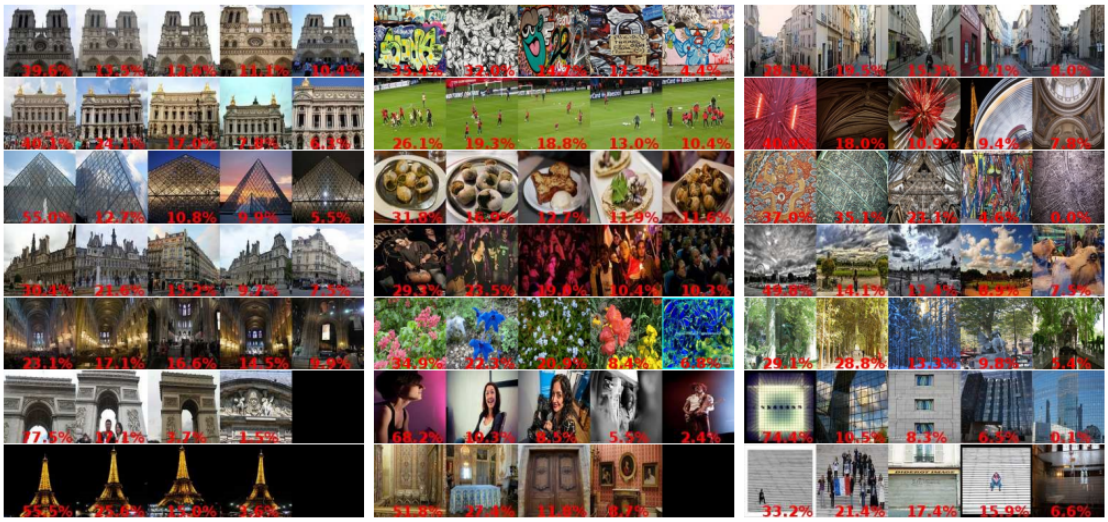
6.3.1. Fast and Robust Archetypal Analysis for Representation Learning

Participants: Yuansi Chen, Julien Mairal, Zaid Harchaoui.

In [9], we revisit a pioneer unsupervised learning technique called archetypal analysis, which is related to successful data analysis methods such as sparse coding and non-negative matrix factorization. Since it was proposed, archetypal analysis did not gain a lot of popularity even though it produces more interpretable models than other alternatives. Because no efficient implementation has ever been made publicly available, its application to important scientific problems may have been severely limited. Our goal is to bring back into favour archetypal analysis. We propose a fast optimization scheme using an active-set strategy, and provide

an efficient open-source implementation interfaced with Matlab, R, and Python. Then, we demonstrate the usefulness of archetypal analysis for computer vision tasks, such as codebook learning, signal classification, and large image collection visualization.

In Figure 5, we present some archetypes corresponding to the request "Paris" when downloading 36 600 images uploaded in 2012 and 2013, and sorted by relevance on the Flickr website.



(a) Archetypes representing (expected) landmarks. (b) Less expected archetypes. (c) Archetypes representing scene composition.

Figure 5. Classical landmarks appear on the left, which is not surprising since Flickr contains a large number of vacation pictures. In the middle, we display several archetypes that we did not expect, including ones about soccer, graffitis, food, flowers, and social gatherings. Finally, we display on the right some archetypes that do not seem to have some semantic meaning, but they capture some scene composition or texture that are common in the dataset.

6.3.2. Conditional Gradient Algorithms for Norm-Regularized Smooth Convex Optimization

Participants: Zaid Harchaoui, Anatoli Juditsky, Arkadii Nemirovski.

In this paper [6], we consider convex optimization problems arising in machine learning in high-dimensional settings. For several important learning problems, such as e.g. noisy matrix completion, state-of-the-art optimization approaches such as composite minimization algorithms are difficult to apply and do not scale up to large datasets. We study three conditional gradient-type algorithms, *i.e.* first-order optimization algorithms that require a linear minimization oracle but do not require a proximal oracle. These new algorithms are suitable for large-scale problems, and enjoy finite-time convergence guarantees. Promising experimental results are presented on two large-scale real-world datasets. The method is illustrated in Figure 6.

6.3.3. A Smoothing Approach for Composite Conditional Gradient with Nonsmooth Loss

Participants: Federico Pierucci, Zaid Harchaoui, Jérôme Malick [BIPOP Team, Inria].

In [25], we consider learning problems where the nonsmoothness lies both in the convex empirical risk and in the regularization penalty. Examples of such problems include learning with nonsmooth loss functions and atomic decomposition regularization penalty. Such doubly nonsmooth learning problems prevent the use of recently proposed composite conditional gradient algorithms for training, which are particularly attractive for large-scale applications. Indeed, they rely on the assumption that the empirical risk part of the objective is smooth. We propose a composite conditional gradient algorithm with smoothing to tackle such learning

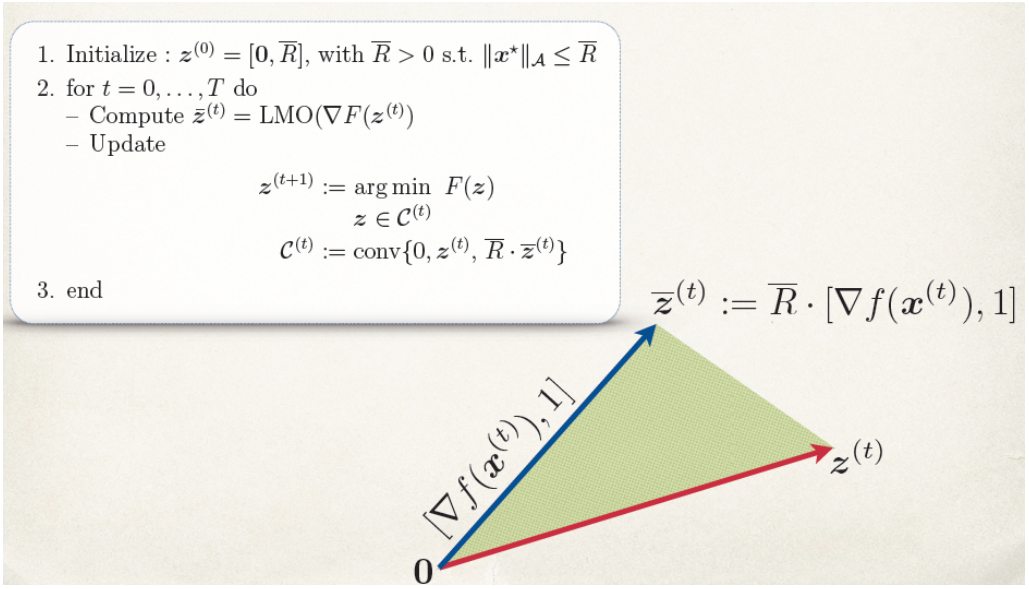


Figure 6. Overview of the composite conditional gradient algorithm which minimizes $F(x) := f(x) + \lambda\|x\|_{\mathcal{A}}$, where f is smooth and $\|\cdot\|_{\mathcal{A}}$ is an atomic-decomposition norm.

problems. We set up a framework allowing to systematically design parametrized smooth surrogates of nonsmooth loss functions. We then propose a smoothed composite conditional gradient algorithm, for which we prove theoretical guarantees on the accuracy. We present promising experimental results on collaborative filtering tasks (see Figure 7).

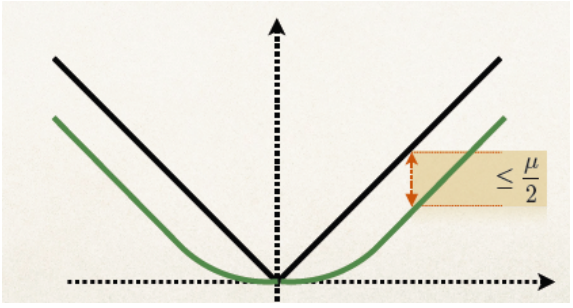


Figure 7. Illustration of the smooth surrogate with parameter μ (green) of the absolute value function (black).

6.3.4. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning

Participant: Julien Mairal.

In this paper [27], we study optimization methods consisting of iteratively minimizing surrogates of an objective function, as illustrated in Figure 8. We introduce a new incremental scheme that experimentally matches or outperforms state-of-the-art solvers for large-scale optimization problems typically arising in machine learning.

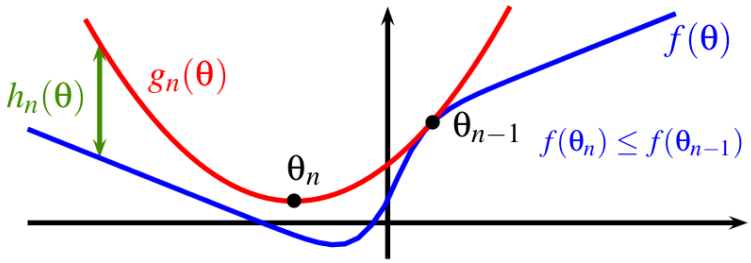


Figure 8. Illustration of the basic majorization-minimization principle. We compute a surrogate g_n of the objective function f around a current estimate θ_{n-1}. The new estimate θ_n is a minimizer of g_n. The approximation error h_n is smooth.

6.3.5. *Efficient RNA Isoform Identification and Quantification from RNA-Seq Data with Network Flows*

Participants: Elsa Bernard [Institut Curie, Ecoles des Mines-ParisTech], Laurent Jacob [CNRS, LBBE Laboratory], Julien Mairal [correspondant], Jean-Philippe Vert [Institut Curie, Ecoles des Mines-ParisTech].

Several state-of-the-art methods for isoform identification and quantification are based on ℓ_1-regularized regression, such as the Lasso. However, explicitly listing the—possibly exponentially—large set of candidate transcripts is intractable for genes with many exons. For this reason, existing approaches using the ℓ_1-penalty are either restricted to genes with few exons or only run the regression algorithm on a small set of preselected isoforms. In [4], we introduce a new technique called FlipFlop, which can efficiently tackle the sparse estimation problem on the full set of candidate isoforms by using network flow optimization. Our technique removes the need of a preselection step, leading to better isoform identification while keeping a low computational cost. Experiments with synthetic and real RNA-Seq data confirm that our approach is more accurate than alternative methods and one of the fastest available. Figure 9 presents the graph on which the network flow optimization is performed.

6.3.6. *Riemannian Sparse Coding for Positive Definite Matrices*

Participants: Anoop Cherian, Suvrit Sra [MPI].

Inspired by the great success of sparse coding for vector valued data, our goal in this work [12] is to represent symmetric positive definite (SPD) data matrices as sparse linear combinations of atoms from a dictionary, where each atom itself is an SPD matrix. Since SPD matrices follow a non-Euclidean (in fact a Riemannian) geometry, existing sparse coding techniques for Euclidean data cannot be directly extended. Prior works have approached this problem by defining a sparse coding loss function using either extrinsic similarity measures (such as the log-Euclidean distance) or kernelized variants of statistical measures (such as the Stein divergence, Jeffrey's divergence, etc.). In contrast, we propose to use the intrinsic Riemannian distance on the manifold of SPD matrices. Our main contribution is a novel mathematical model for sparse coding of SPD matrices; we also present a computationally simple algorithm for optimizing our model. Experiments on several computer vision datasets showcase superior classification and retrieval performance compared against state-of-the-art approaches.

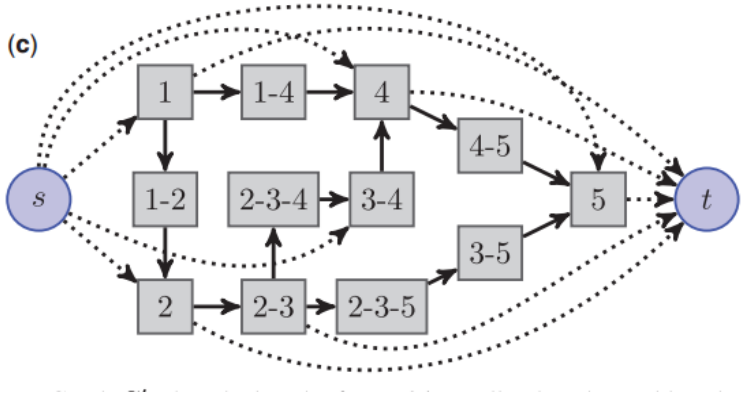


Figure 9. Graph on which we perform network flow optimization. Nodes represent observed reads, and paths on the graph correspond to isoforms.

6.4. Recognition in video

6.4.1. Occlusion and Motion Reasoning for Long-Term Tracking

Participants: Yang Hua, Karteek Alahari, Cordelia Schmid.

Object tracking is a reoccurring problem in computer vision. Tracking-by-detection approaches, in particular Struck, have shown to be competitive in recent evaluations. However, such approaches fail in the presence of long-term occlusions as well as severe viewpoint changes of the object. In this paper we propose a principled way to combine occlusion and motion reasoning with a tracking-by-detection approach. Occlusion and motion reasoning is based on state-of-the-art long-term trajectories which are labeled as object or background tracks with an energy-based formulation. The overlap between labeled tracks and detected regions allows to identify occlusions. The motion changes of the object between consecutive frames can be estimated robustly from the geometric relation between object trajectories. If this geometric change is significant, an additional detector is trained. Experimental results show that our tracker obtains state-of-the-art results and handles occlusion and viewpoints changes better than competing tracking methods. This work corresponds to the publication [15] and is illustrated in Figure 10.

6.4.2. Category-Specific Video Summarization

Participants: Danila Potapov, Matthijs Douze, Zaid Harchaoui, Cordelia Schmid.

In large video collections with clusters of typical categories, such as “birthday party” or “flash-mob”, category-specific video summarization can produce higher quality video summaries than unsupervised approaches that are blind to the video category. Given a video from a known category, our approach published in [20] first efficiently performs a temporal segmentation into semantically-consistent segments, delimited not only by shot boundaries but also general change points. Then, equipped with an SVM classifier, our approach assigns importance scores to each segment. The resulting video assembles the sequence of segments with the highest scores, as shown in Figure 11. The obtained video summary is therefore both short and highly informative. Experimental results on videos from the multimedia event detection (MED) dataset of TRECVID’11 show that our approach produces video summaries with higher relevance than the state of the art.

6.4.3. Efficient Action Localization with Approximately Normalized Fisher Vectors

Participants: Dan Oneata, Jakob Verbeek, Cordelia Schmid.

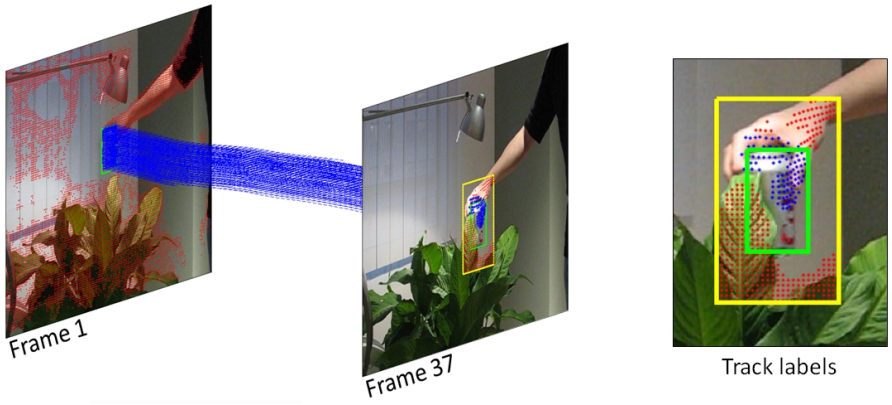


Figure 10. Left: Long-term tracks beginning in frame 1 of the Coke sequence. The yellow box shows the search region used to compute the bounding box most likely to contain the object (green box). We use the tracks to estimate the object state. Right: Close-up of the track labels in frame 37. Here, less than 60% of the tracks within the predicted bounding box are assigned to the object (blue), and the remaining are labelled as background (red). Thus, the object is predicted to be in an occluded state.

Figure 11. Original video, and its video summary for the category "birthday party".

The Fisher vector (FV) representation is a high-dimensional extension of the popular bag-of-word representation. Transformation of the FV by power and ℓ_2 normalizations has shown to significantly improve its performance, and led to state-of-the-art results for a range of image and video classification and retrieval tasks. These normalizations, however, render the representation non-additive over local descriptors. Combined with its high dimensionality, this makes the FV computationally expensive for the purpose of localization tasks. In [18] we present approximations to both these normalizations (see Figure 12), which yield significant improvements in the memory and computational costs of the FV when used for localization. Second, we show how these approximations can be used to define upper-bounds on the score function that can be efficiently evaluated, which enables the use of branch-and-bound search as an alternative to exhaustive sliding window search. We present experimental evaluation results on classification and temporal localization of actions in videos. These show that the our approximations lead to a speedup of at least one order of magnitude, while maintaining state-of-the-art action recognition and localization performance.

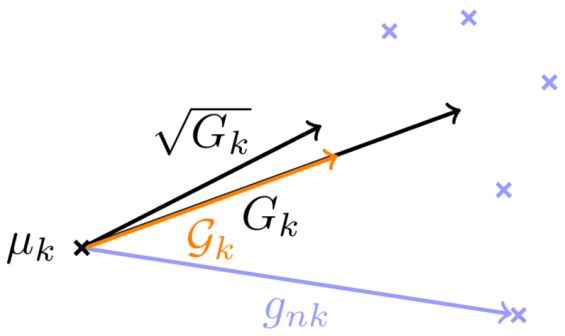


Figure 12. Schematic illustration of the proposed approximation for the square-root normalization. We depict a Fisher vector G_k as an aggregation of individual gradients g_{nk}. Both the exact ($\sqrt{G_k}$) and the approximated ($\mathcal{G}_k$) square-root normalizations scale similarly the Fisher vector G_k; the approximated variant has the property of preserving the orientation of the Fisher vector G_k.

6.4.4. Spatio-Temporal Object Detection Proposals

Participants: Dan Oneata, Jakob Verbeek, Cordelia Schmid, Jerome Revaud.

Spatio-temporal detection of actions and events in video is a challenging problem. Besides the difficulties related to recognition, a major challenge for detection in video is the size of the search space defined by spatio-temporal tubes formed by sequences of bounding boxes along the frames. Recently methods that generate unsupervised detection proposals have proven to be very effective for object detection in still images. These methods open the possibility to use strong but computationally expensive features since only a relatively small number of detection hypotheses need to be assessed. In [17] we make two contributions towards exploiting detection proposals for spatio-temporal detection problems. First, we extend a recent 2D object proposal method, to produce spatio-temporal proposals by a randomized supervoxel merging process (see Figure 13). We introduce spatial, temporal, and spatio-temporal pairwise supervoxel features that are used to guide the merging process. Second, we propose a new efficient supervoxel method. We experimentally evaluate our detection proposals, in combination with our new supervoxel method as well as existing ones. This evaluation shows that our supervoxels lead to more accurate proposals when compared to using existing state-of-the-art supervoxel methods.

6.4.5. EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow

Participants: Revaud Jerome, Weinzaepfel Philippe, Harchaoui Zaid, Cordelia Schmid.

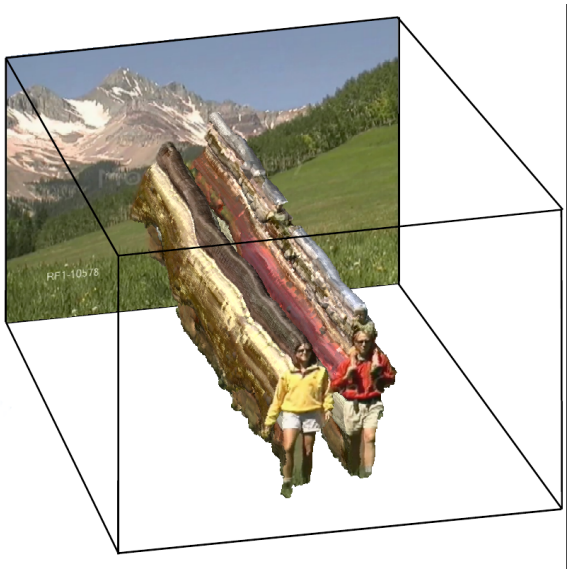


Figure 13. An example of a spatio-temporal proposal generated by our randomized supervoxel merging process. The video sample belongs to the UCF Sports dataset.

We propose a novel approach [29] for optical flow estimation, targeted at large displacements with significant occlusions. It consists of two steps: i) dense matching by edge-preserving interpolation from a sparse set of matches; ii) variational energy minimization initialized with the dense matches. The sparse-to-dense interpolation relies on an appropriate choice of the distance, namely an edge-aware geodesic distance. This distance is tailored to handle occlusions and motion boundaries (see Figure 14), two common and difficult issues for optical flow computation. We also propose an approximation scheme for the geodesic distance to allow fast computation without loss of performance. Subsequent to the dense interpolation step, standard one-level variational energy minimization is carried out on the dense matches to obtain the final flow estimation. The proposed approach, called Edge-Preserving Interpolation of Correspondences (EpicFlow) is fast and robust to large displacements. It significantly outperforms the state of the art on MPI-Sintel and performs on par on KITTI and Middlebury.

6.4.6. Weakly Supervised Action Labeling in Videos Under Ordering Constraints.

Participants: Piotr Bojanowski [Willow team, Inria], Rémi Lajugie [Willow team, Inria], Francis Bach [Sierra team, Inria], Ivan Laptev [Willow team, Inria], Jean Ponce [Willow team, Inria], Cordelia Schmid, Josef Sivic [Willow team, Inria].

Suppose we are given a set of video clips, each one annotated with an ordered list of actions, such as “walk” then “sit” then “answer phone” extracted from, for example, the associated text script. See Fig. 15 for an illustration. In this work [8], we seek to temporally localize the individual actions in each clip as well as to learn a discriminative classifier for each action. We formulate the problem as a weakly supervised temporal assignment with ordering constraints. Each video clip is divided into small time intervals and each time interval of each video clip is assigned one action label, while respecting the order in which the action labels appear in the given annotations. We show that the action label assignment can be determined together with learning a classifier for each action in a discriminative manner. We evaluate the proposed model on a new and challenging dataset of 937 video clips with a total of 787720 frames containing sequences of 16 different actions from 69 Hollywood movies.

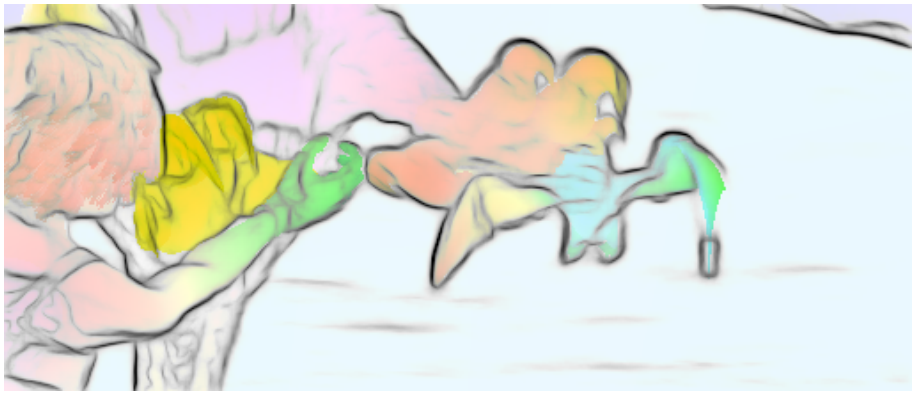


Figure 14. Image edges detected with SED and ground-truth optical flow. Motion discontinuities appear most of the time at image edges

Figure 15. Sample data used as input to our method. Every video clip comes with an ordered list of actions that appears in it. These actions are not temporaly localized, only the order is known. The goal of our paper is to correctly localize these actions according to a discriminative criterion.

6.4.7. Mixing Body-Part Sequences for Human Pose Estimation

Participants: Cherian Anoop, Mairal Julien, Alahari Karteek, Schmid Cordelia.

This work [11] presents a method for estimating articulated human poses in videos. We cast this as an optimization problem defined on body parts with spatio-temporal links between them. The resulting formulation is unfortunately intractable and previous approaches only provide approximate solutions. Although such methods perform well on certain body parts, e.g., head, their performance on lower arms, i.e., elbows and wrists, remains poor. We present a new approximate scheme with two steps dedicated to pose estimation. First, our approach takes into account temporal links with subsequent frames for the less-certain parts, namely elbows and wrists. Second, our method decomposes poses into limbs, generates limb sequences across time, and recomposes poses by mixing these body part sequences (See Figure 16 for an illustration). We introduce a new dataset "Poses in the Wild", which is more challenging than the existing ones, with sequences containing background clutter, occlusions, and severe camera motion. We experimentally compare our method with recent approaches on this new dataset as well as on two other benchmark datasets, and show significant improvement.

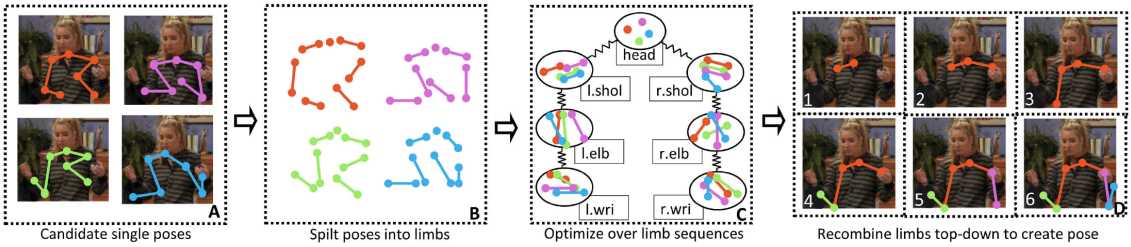


Figure 16. Illustration of our limb recombination scheme. From left to right: Block-A: An image and four candidate poses, where only a part of each pose is well-aligned with the person. Block-B: We divide each candidate pose into limb parts. Block-C: We allow the recombination of limbs from different pose candidates with constraints between two limbs that have a joint in common. Block-D: An example where recombination builds an accurate pose, which is not in the original candidate set.

6.4.8. The LEAR Submission at Thumos 2014

Participants: Dan Oneata, Jakob Verbeek, Cordelia Schmid.

In [28] we describe the submission of our team to the THUMOS workshop in conjunction with ECCV 2014. Our system is based on Fisher vector (FV) encoding of dense trajectory features (DTF), which we also used in our 2013 submission. The dataset is based on the UCF101 dataset, which is currently the largest action dataset both in terms of number of categories and clips, with more than 13000 clips drawn from 101 action classes. This year special attention was paid to classification of uncropped videos, where the action of interest appears in videos that contain also non-relevant sections. This year's submission additionally incorporated static-image features (SIFT, Color, and CNN) and audio features (ASR and MFCC) for the classification task. For the detection task, we combined scores from the classification task with FV-DTF features extracted from video slices. We found that these additional visual and audio feature significantly improve the classification results. For localization we found that using the classification scores as a contextual feature besides local motion features leads to significant improvements. In Figure 17 we show the middle frame from the top four ranked videos corresponding to the three hardest classes (as evaluated on the validation data). Our team has ranked second on the classification challenge (out of eleven teams) and first on the detection challenge (out of three teams).

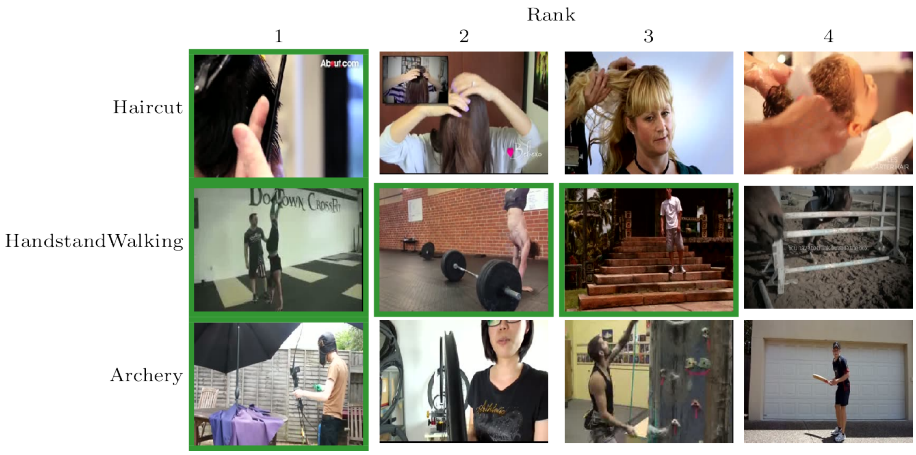


Figure 17. Snapshots from the top four ranked test videos for the three hardest classes; green borders indicate true positives.

6.4.9. The LEAR Submission at TrecVid MED 2014

Participants: Matthijs Douze, Dan Oneata, Mattis Paulin, Clément Leray, Nicolas Chesneau, Danila Potapov, Jakob Verbeek, Karteek Alahari, Zaid Harchaoui, Lori Lamel [Spoken Language Processing group, LIMSI, CNRS], Jean-Luc Gauvain [Spoken Language Processing group, LIMSI, CNRS], Christoph Schmidt [Fraunhofer IAIS, Sankt Augustin], Cordelia Schmid.

In [26] we describe our participation to the 2014 edition of the TrecVid Multimedia Event Detection task. Our system is based on a collection of local visual and audio descriptors, which are aggregated to global descriptors, one for each type of low-level descriptor, using Fisher vectors. Besides these features, we use two features based on convolutional networks: one for the visual channel, and one for the audio channel. Additional high-level features are extracted using ASR and OCR features. Finally, we used mid-level attribute features based on object and action detectors trained on external datasets. In the notebook paper we present an overview of the features and the classification techniques, and experimentally evaluate our system on TrecVid MED 2011 data.

We participated in four tasks, which differ in the amount of training videos for each event (either 10 or 100), and the time that is allowed for the processing. For the 20 pre-specified events several weeks are allowed to extract features, train models, and to score the test videos (which consisted of 8,000 hours of video this year). For the 10 ad-hoc events, we only have five days to do all processing. Across the 11 participating teams, our results ranked first for the 10-example ad-hoc task, and fourth and fifth place for the other tasks.

7. Bilateral Contracts and Grants with Industry

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

Participants: Anoop Cherian, Zaid Harchaoui, Yang Hua, Cordelia Schmid, Karteek Alahari.

This collaborative project, which started in September 2008, brings together the WILLOW and LEAR 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. Yang Hua is funded by this project.

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

Participants: Julien Mairal, Zaid Harchaoui.

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.

7.3. WayWay, OMB LABS

Participants: Matthijs Douze, Julien Mairal, Mattis Paulin, Jerome Revaud, Cordelia Schmid.

The collaboration with OMB Labs consisted of transferring technology developed at LEAR for large-scale image classification for the web application wayway.us. The company is developing a smartphone application for recommending restaurants and social places in US cities by exploiting image content from Instagram. Their system requires automatically classifying Instagram images into a few well-defined categories, “food", “people" and “atmosphere". Through a consulting project, with visits of engineers from OMB Labs, the team has helped them develop a full image classification pipeline to suit their industrial needs.

7.4. Xerox Research Center Europe

Participants: Matthijs Douze, Zaid Harchaoui, Mattis Paulin, Cordelia Schmid.

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 large-scale recognition and deep learning based image description.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR Project Qcompere

Participants: Guillaume Fortier, Cordelia Schmid, Jakob Verbeek.

This three-and-a-half year project started in November 2010. It is aimed at identifying people in video using both audio (using speech and speaker recognition) and visual data in challenging footage such as news broadcasts, or movies. The partners of this project are the CNRS laboratories LIMSI and LIG, the university of Caen, Inria’s LEAR team, as well as two industrial partners Yacast and Vecsys Research.

8.1.2. ANR Project Physionomie

Participants: Frédéric Jurie [University of Caen], Jakob Verbeek, Shreyas Saxena.

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 LEAR.

8.1.3. ANR Project Macaron

Participants: Julien Mairal, Zaid Harchaoui, 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.

8.1.4. PEPS CNRS BMI (Biology - Mathematics - Computer Science), Project FlipFlop

Participants: Elsa Bernard [Institut Curie, Ecoles des Mines-ParisTech], Laurent Jacob [CNRS, LBBE Laboratory], Julien Mairal, Jean-Philippe Vert [Institut Curie, Ecoles des Mines-ParisTech], Anne-Hélène Monsoro-Burq [Institut Curie].

The project is concerned with large-scale sparse estimation techniques for processing RNA-Seq data. It led to a joint publication [4] with partners from Inria Grenoble, Institut Curie in Paris, and the LBBE laboratory in Lyon. The principal investigator was Laurent Jacob (CNRS, LBBE laboratory). The project started in Jun 2012 and ended in Dec 2014.

8.1.5. MASTODONS Program CNRS - Project Gargantua

Participants: Zaid Harchaoui, Julien Mairal.

The project is concerned with machine learning and mathematical optimization for big data. The partners are from LJK (Grenoble), LIG (Grenoble), LIENS (ENS, Paris), Lab. P. Painleve (Lille). Principal investigator/leader: Zaid Harchaoui. Dates: May 2013-Dec. 2014

8.1.6. Equipe-action ADM du Labex Persyval (Grenoble) "Khronos"

Participants: Zaid Harchaoui, Massih-Reza Amini [LIG].

The partners of this project are from the laboratories LJK, LIG, GIPSA, TIMC, CEA. The principal investigators/leaders are Zaid Harchaoui (Inria and LJK), Massih-Reza Amini (LIG). The project started in Jan. 2014 and ends in Dec. 2016.

8.2. European Initiatives

8.2.1. AXES

Participants: Ramazan Cinbis, Matthijs Douze, Zaid Harchaoui, Dan Oneata, Danila Potapov, Cordelia Schmid, Jakob Verbeek, Clement Leray, Anoop Cherian.

This 4-year project started in January 2011 and ends in March 2015. Its goal is to develop and evaluate tools to analyze and navigate large video archives, eg. from broadcasting services. The partners of the project are ERCIM, Univ. of Leuven, Univ. of Oxford, LEAR, Dublin City Univ., Fraunhofer Institute, Univ. of Twente, BBC, Netherlands Institute of Sound and Vision, Deutsche Welle, Technicolor, EADS, Univ. of Rotterdam. See http://www.axes-project.eu/ for more information.

8.2.2. ERC Advanced grant Allegro

Participants: Cordelia Schmid, Karteek Alahari, Jerome Revaud, Pavel Tokmakov, Nicolas Chesneau.

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.

8.3. International Initiatives

8.3.1. Inria Associate Teams

- **HYPERION: Large-scale statistical learning for visual recognition:** Zaid Harchaoui and Cordelia Schmid have an ongoing collaboration resp. with Pr. Jitendra Malik (EECS) and Pr. Nourredine El Karoui (Stat. dpt.) of UC Berkeley in the fall 2011. This collaboration has been supported by the *associated team "Hyperion"* and the *France-Berkeley Fund* (dates: June 2012-Dec. 2013). The collaboration is focusing on *large-scale statistical learning for computer vision*, ranging from the high-dimensional statistics aspects to real-world applications on large image and video datasets. Several visits of members of each institution and co-supervision of students happened in 2012, 2013, 2014. As part of the "Hyperion" associated team, two papers were published resp. in CVPR'14 and ICML'14, and one paper is currently in revision.

8.3.2. Inria International Partners

- **UC Berkeley:** This collaboration between Bin Yu, Jack Gallant, Yuval Benjamini, Adam Bloniarz (UC Berkeley), Ben Willmore (Oxford University) and Julien Mairal (Inria LEAR) 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. We are planning to welcome one student from UC Berkeley during the summer 2015 to work on this project.
- **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.
- **MPI Tübingen:** C. Schmid collaborates with M. Black, a research director at MPI. In 2013, she spent one month at MPI and worked with a PhD student, S. Zuffi, and a postdoctoral researcher, H. Jhuang. C. Schmid has continued this collaboration in 2014 and spent also one month there.

8.3.3. Participation in Other International Programs

- **France-Berkeley fund:** The LEAR team was awarded in 2014 a grant from the France-Berkeley fund for a project between Julien Mairal and 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 for a period of one year, from November 2014 to November 2015. The funds are meant to support scientific and scholarly exchanges and collaboration between the two teams.

8.4. International Research Visitors

8.4.1. Visits to International Teams

- **Sabbatical program** Zaid Harchaoui is currently on sabbatical at New-York university, from October 2014 to September 2015.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. General chair, scientific chair

- C. Schmid is general chair for IEEE Conference on Computer Vision and Pattern Recognition, 2015.

9.1.1.2. Member of the organizing committee

- C. Schmid and K. Alahari organized the ALLEGRO Workshop on weakly supervised learning and video recognition, Grenoble, 2014.
- Z. Harchaoui organized the "Tutorial on Frank-Wolfe and Greedy Optimization for Learning with Big Data" at ICML 2014, June 2014.
- Z. Harchaoui has organized a summer-school "Khronos Days", focused on "High-Dimensional Learning and Optimization", as part of the "Khronos" project (LabEx Persyval-Lab) and the "Gargantua" project (CNRS-Mastodons). http://lear.inrialpes.fr/people/harchaoui/projects/khronos/.
- Z. Harchaoui organized one international workshop. "Optimization and Statistical Learning workshop", Les Houches, France, January 2015. http://lear.inrialpes.fr/workshop/osl2015/.
- Z. Harchaoui co-organized one national GDR-ISIS workshop on "Learning representations and signal processing".
- Z. Harchaoui co-organized NIPS'14 "Optimization for Machine Learning" Workshop in Montreal (Canada).
- J. Mairal organized the session "Optimization and statistics" at "Journées MAS", Toulouse, France.
- F. Pierucci co-organized the Grenoble Optimization Day workshop.

9.1.2. Scientific events selection

9.1.2.1. Member of the conference program committee

- J. Verbeek: area chair for BMVC '14, ECCV '14, CVPR '15.
- Z. Harchaoui: area chair for ICML '15.
- J. Mairal: area chair for ICML '15.

9.1.2.2. 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.

9.1.3. Journal

9.1.3.1. Member of the editorial board

- 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: Guest editor for the Special Issue on Sparse Coding of the International Journal of Computer Vision.
- J. Mairal: Associate editor for IEEE Signal Processing Letters, since August 2014.
- K. Alahari: Co-guest editor: Special Issue on "Higher Order Graphical Models in Computer Vision: Modelling, Inference & Learning", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015 (in press)

9.1.3.2. Reviewer

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), image processing (SIAM Imaging Science), information theory (IEEE Transactions on Information Theory), statistics (Bernoulli, Annals of Statistics, Journal of American Statistical Association).

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Doctorat: Z. Harchaoui, Lecture *Large-scale machine learning*, in “Tutorial on Large-Scale Visual Recognition” at CVPR 2014.

Doctorat: K. Alahari, Half-day Tutorial on MAP Estimation and Structured Prediction at ICPR, Stockholm, August 2014.

Doctorat: J. Mairal, course “Sparse estimation for image and vision processing”, 9H eqTD, DENIS summer school at Tampere university, Finland;

Doctorat: J. Mairal, course “Sparse estimation for image and vision processing”, 4.5H eqTD, given at PSL research university, Paris, France.

Master : M. Douze and K. Alahari. Multimedia databases, 3rd year ENSIMAG, Grenoble, France.

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

Master: J. Mairal. “Statistical Learning and Applications”. 13.5H eqTD, M2, Ecole Normale Supérieure de Lyon, France.

Master : C. Schmid, “Object recognition and computer vision”, 10H, M2, ENS Cachan, France.

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

Master: P. Weinzaepfel, “Réseaux IP”, 18h TD, M1, University Joseph Fourier, Grenoble, France.

Master: P. Weinzaepfel, “Introduction aux Réseaux”, 15H TD, M1, University Joseph Fourier, Grenoble, France.

Licence : F. Pierucci, “Apprentissage du raisonnement, algèbre linéaire et analyse élémentaire.”. 38H, L1, Grenoble univ., France.

Licence : H. Lin, “Apprentissage du raisonnement, algèbre linéaire et analyse élémentaire.”. 38H, L1, Grenoble univ., France.

Licence: P. Weinzaepfel, “Introduction à UNIX et à la programmation en langage C”, 33.5H TD, L1, DLST Grenoble, France.

9.2.2. Supervision

PhD: G. Cinbis, “Fisher kernel based models for image classification and object localization”, Grenoble Univ, July 2014, advisors: C. Schmid and J. Verbeek.

PhD: Z. Akata, “"Contributions to Large-Scale Learning for Image Classification”, Grenoble Univ, Januray 2014, advisors: C. Schmid and F. Perronnin.

9.2.3. Juries

- C. Schmid, PhD committe for Mihir Jain, Inria Rennes, April 2014.
- C. Schmid, PhD committe for Mingyuan JIU, INSA Lyon, April 2014.
- C. Schmid, examiner for HdR Josef Sivic, ENS, Feburary 2014.
- J. Mairal, PhD committee for Nicolas Duforet-Frebourg, Grenoble University, October 2014.
- K. Alahari: President of the PhD committee for Jon Almazan, Computer Vision Center, Barcelona, Spain, October 2014.
- K. Alahari: PhD committee for Heydar Maboudi, KTH Royal Institute of Technology, Stockholm, Sweden, May 2014.

9.3. Other responsabilities

- C. Schmid was member of the evaluation panel for ERC starting grants, 2014.
- C. Schmid is member of the PAMI-TC awards committee, and the PAMI-TC executive commitee.
- J. Verbeek: reviewer for grant proposals for the Indo French Centre for the Promotion of Advanced Research (CEFIPRA).
- Z. Harchaoui is in the scientific board of area “Machine Learning” of GDR ISIS (Groupe de Recherche “Information, Signal, Image et ViSion”).
- J. Mairal: reviewer for grant proposals from ANR and ERC.

9.4. Invited presentations

9.4.1. Keynote talks

- C. Schmid: Keynote speaker at Annual Workshop of the Austrian Association for Pattern Recognition, IST Austria, May 2014.
- C. Schmid: Keynote speaker at Netherlands Conference on Computer Vision, April 2014.

9.4.2. Invited talks

- Z. Harchaoui: Seminar, Microsoft Research, New York, April 2014.
- Z. Harchaoui: Seminar, Newton Institute, Cambridge, January 2014.
- Z. Harchaoui: Seminar, CBLL, NYU, April 2014.
- C. Schmid: Invited speaker at First French-German Mathematical Image Analysis Conference, Paris, January 2014.
- C. Schmid: Seminar at Univ. Edinburgh, July 2014.
- C. Schmid: Seminar at MPI, Tübingen, April 2014.
- J. Verbeek: Invited speaker at the Croatian Computer Vision Workshop, Zagreb, 2014.
- J. Mairal: invited speaker in mini-symposium at SIAM conference on Optimization, San Diego, 2014.
- J. Mairal: invited speaker at Journées BIG Data, Toulouse, 2014.

- J. Mairal: invited speaker at Journées MAS, Toulouse, 2014.
- K. Alahari: Seminar, Computer Vision Center (CVC), Barcelona, Spain, October 2014.
- K. Alahari: Seminar, University of California, Berkeley, USA, July 2014.
- K. Alahari: Invited speaker, Tutorial on Learning and Inference in Discrete Graphical Models, CVPR, Columbus, USA, June 2014.
- K. Alahari: Seminar, KTH Royal Institute of Technology, Stockholm, Sweden, May 2014.
- A. Cherian: Seminar at Yahoo Labs, Bangalore, India, April 2014.
- A. Cherian: Seminar at Siemens Corporate Research, Princeton, NJ, USA, June 2014.
- P. Weinzapfael: Seminar at UC Berkeley, December, 2014.

9.5. Popularization

- Zaid Harchaoui and Martin Jaggi published in the Mathematical Optimization Society newsletter "Optima" an interview of Marguerite Frank, co-inventor of the Frank-Wolfe algorithm for constrained optimization, with a companion historical survey.
- Julien Mairal, Francis Bach, and Jean Ponce have published a monograph [24] "Sparse Modeling for Image and Vision Processing", to appear in Foundations and Trends in Computer Graphics and Vision.
- Together with other students in LJK, Mattis Paulin did an interview of Emmanuel Candes for his award of the Prix Jean Kuntzmann in June 2014. The interview will soon be published in Journal du CNRS.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] Z. AKATA. *Contributions to large-scale learning for image classification*, Université de Grenoble, January 2014, https://tel.archives-ouvertes.fr/tel-00873807

[2] R. G. CINBIS, C. SCHMID, J. VERBEEK. *Fisher kernel based models for image classification and object localization*, Université de Grenoble, July 2014, https://tel.archives-ouvertes.fr/tel-01071581

Articles in International Peer-Reviewed Journals

[3] Z. AKATA, F. PERRONNIN, Z. HARCHAOUI, C. SCHMID. *Good Practice in Large-Scale Learning for Image Classification*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", March 2014, vol. 36, n^{o} 3, pp. 507-520 [*DOI :* 10.1109/TPAMI.2013.146], https://hal.inria.fr/hal-00835810

[4] E. BERNARD, L. JACOB, J. MAIRAL, J.-P. VERT. *Efficient RNA Isoform Identification and Quantification from RNA-Seq Data with Network Flows*, in "Bioinformatics", September 2014, vol. 30, n^{o} 17, pp. 2447-2455 [*DOI :* 10.1093/BIOINFORMATICS/BTU317], https://hal-mines-paristech.archives-ouvertes.fr/hal-00803134

[5] A. GAIDON, Z. HARCHAOUI, C. SCHMID. *Activity representation with motion hierarchies*, in "International Journal of Computer Vision", May 2014, vol. 107, n^{o} 3, pp. 219-238 [*DOI :* 10.1007/s11263-013-0677-1], https://hal.inria.fr/hal-00908581

[6] Z. HARCHAOUI, A. JUDITSKY, A. S. NEMIROVSKI. *Conditional Gradient Algorithms for Norm-Regularized Smooth Convex Optimization*, in "Mathematical Programming, Series A", April 2014, pp. 1-30, 30 pages [*DOI :* 10.1007/s10107-014-0778-9], https://hal.archives-ouvertes.fr/hal-00978368

[7] G. SEGUIN, K. ALAHARI, J. SIVIC, I. LAPTEV. *Pose Estimation and Segmentation of Multiple People in Stereoscopic Movies*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2014, 1 p. [*DOI :* 10.1109/TPAMI.2014.2369050], https://hal.inria.fr/hal-01089660

International Conferences with Proceedings

[8] P. BOJANOWSKI, R. LAJUGIE, F. BACH, I. LAPTEV, J. PONCE, C. SCHMID, J. SIVIC. *Weakly Supervised Action Labeling in Videos Under Ordering Constraints*, in "ECCV - European Conference on Computer Vision", Zurich, Switzerland, September 2014, pp. 628-643 [*DOI :* 10.1007/978-3-319-10602-1_41], https://hal.inria.fr/hal-01053967

[9] Y. CHEN, J. MAIRAL, Z. HARCHAOUI. *Fast and Robust Archetypal Analysis for Representation Learning*, in "CVPR 2014 - IEEE Conference on Computer Vision & Pattern Recognition", Columbus, United States, June 2014, https://hal.inria.fr/hal-00995911

[10] A. CHERIAN. *Nearest Neighbors Using Compact Sparse Codes*, in "ICML - 31st International Conference on Machine Learning", Beijing, China, June 2014, https://hal.inria.fr/hal-01057690

[11] A. CHERIAN, J. MAIRAL, K. ALAHARI, C. SCHMID. *Mixing Body-Part Sequences for Human Pose Estimation*, in "CVPR 2014 - IEEE Conference on Computer Vision & Pattern Recognition", Columbus, OH, United States, IEEE, June 2014, https://hal.inria.fr/hal-00978643

[12] A. CHERIAN, S. SRA. *Riemannian Sparse Coding for Positive Definite Matrices*, in "ECCV 2014 - European Conference on Computer Vision", Zurich, Switzerland, September 2014, https://hal.inria.fr/hal-01057703

[13] R. G. CINBIS, J. VERBEEK, C. SCHMID. *Multi-fold MIL Training for Weakly Supervised Object Localization*, in "CVPR 2014 - IEEE Conference on Computer Vision & Pattern Recognition", Columbus, United States, IEEE, June 2014, https://hal.inria.fr/hal-00975746

[14] M. DOUZE, H. JÉGOU. *The Yael library*, in "ACM Multimedia", Orlondo, United States, November 2014, https://hal.inria.fr/hal-01020695

[15] Y. HUA, K. ALAHARI, C. SCHMID. *Occlusion and Motion Reasoning for Long-term Tracking*, in "ECCV 2014 - European Conference on Computer Vision", Zurich, Switzerland, Springer, September 2014, https://hal.inria.fr/hal-01020149

[16] J. MAIRAL, P. KONIUSZ, Z. HARCHAOUI, C. SCHMID. *Convolutional Kernel Networks*, in "Advances in Neural Information Processing Systems (NIPS)", Montreal, Canada, December 2014, https://hal.inria.fr/hal-01005489

[17] D. ONEATA, J. REVAUD, J. VERBEEK, C. SCHMID. *Spatio-Temporal Object Detection Proposals*, in "ECCV 2014 - European Conference on Computer Vision", Zurich, Switzerland, D. FLEET, T. PAJDLA, B. SCHIELE, T. TUYTELAARS (editors), Springer, September 2014, vol. 8691, pp. 737-752 [*DOI :* 10.1007/978-3-319-10578-9_48], https://hal.inria.fr/hal-01021902

[18] D. ONEATA, J. VERBEEK, C. SCHMID. *Efficient Action Localization with Approximately Normalized Fisher Vectors*, in "CVPR 2014 - IEEE Conference on Computer Vision & Pattern Recognition", Columbus, OH, United States, IEEE, June 2014, https://hal.inria.fr/hal-00979594

[19] M. PAULIN, J. REVAUD, Z. HARCHAOUI, F. PERRONNIN, C. SCHMID. *Transformation Pursuit for Image Classification*, in "CVPR 2014 - IEEE Conference on Computer Vision & Pattern Recognition", Columbus, United States, IEEE, June 2014, https://hal.inria.fr/hal-00979464

[20] D. POTAPOV, M. DOUZE, Z. HARCHAOUI, C. SCHMID. *Category-specific video summarization*, in "ECCV 2014 - European Conference on Computer Vision", Zurich, Switzerland, Springer, September 2014, https://hal.inria.fr/hal-01022967

[21] U. ROY, A. MISHRA, K. ALAHARI, C. JAWAHAR. *Scene Text Recognition and Retrieval for Large Lexicons*, in "ACCV - Asian Conference on Computer Vision", Singapore, Singapore, November 2014, https://hal.inria.fr/hal-01088739

[22] H. O. SONG, R. GIRSHICK, S. JEGELKA, J. MAIRAL, Z. HARCHAOUI, T. DARRELL. *On learning to localize objects with minimal supervision*, in "ICML - 31st International Conference on Machine Learning", Beijing, China, JMLR: W&CP, June 2014, vol. 32, https://hal.inria.fr/hal-00996849

National Conferences with Proceedings

[23] M. PAULIN, J. REVAUD, Z. HARCHAOUI, F. PERRONNIN, C. SCHMID. *Selection itérative de transformations pour la classification d'images*, in "RFIA 2014 - Reconnaissance de Formes et Intelligence Artificielle", Rouen, France, June 2014, https://hal.archives-ouvertes.fr/hal-00988820

Scientific Books (or Scientific Book chapters)

[24] J. MAIRAL, F. BACH, J. PONCE. *Sparse Modeling for Image and Vision Processing*, Foundations and Trends in Computer Graphics and Vision, now publishers, December 2014, vol. 8, n^{o} 2-3, 216 p. [*DOI :* 10.1561/9781680830095], https://hal.inria.fr/hal-01081139

Research Reports

[25] F. PIERUCCI, Z. HARCHAOUI, J. MALICK. *A smoothing approach for composite conditional gradient with nonsmooth loss*, Inria Grenoble, July 2014, n^{o} RR-8662, https://hal.inria.fr/hal-01096630

Other Publications

[26] M. DOUZE, D. ONEATA, M. PAULIN, C. LERAY, N. CHESNEAU, D. POTAPOV, J. VERBEEK, K. ALAHARI, C. SCHMID, L. LAMEL, J.-L. GAUVAIN, C. A. SCHMIDT, Z. HARCHAOUI. *The Inria-LIM-VocR and AXES submissions to Trecvid 2014 Multimedia Event Detection*, 2014, https://hal.inria.fr/hal-01089916

[27] J. MAIRAL. *Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning*, October 2014, https://hal.inria.fr/hal-00948338

[28] D. ONEATA, J. VERBEEK, C. SCHMID. *The LEAR submission at Thumos 2014*, 2014, https://hal.inria.fr/hal-01074442

[29] J. REVAUD, P. WEINZAEPFEL, Z. HARCHAOUI, C. SCHMID. *EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow*, January 2015, https://hal.inria.fr/hal-01097477