
IN PARTNERSHIP WITH:
**CNRS**

**Institut polytechnique de Grenoble**

**Université Joseph Fourier (Grenoble)**

Activity Report 2013

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

**Engineers**

Guillaume Fortier [Inria, OSEO Anvar, until Jun 2013]
Clement Leray [Inria, AXES, from Sep 2013]
Jerome Revaud [Inria, Quaero and European Research Council, from Jun 2011 until Nov 2014]
Franck Thollard [Inria, MBDA, until Sep 2013]
Matthijs Douze [Inria, SED 40 %]

**PhD Students**

Zeynep Akata [Univ. Grenoble, CIFRE grant XRCE, from Jan 2011 until Dec 2013]
Ramazan Cinbis [Univ. Grenoble, from Oct 2010 until March 2014]
Dan Oneata [Univ. Grenoble, FP7 AXES project, from Oct 2011 until Oct 2014]
Vicky Kalogeiton [Univ. Edinburgh, European Research Council, co-supervision with V. Ferrari, from Sep 2013 until Dec 2016]
Mattis Paulin [Univ. Grenoble, from Apr 2013 until Apr 2016]
Federico Pierucci [Univ. Grenoble I, from Jan 2012 until Sep 2015]
Danila Potapov [Univ. Grenoble, FP7 AXES project and Quaero, from Sep 2011 until Aug 2014]
Shreyas Saxena [Univ. Grenoble, ANR PHYSIONOMIE project, from Feb 2013]
Yang Hua [Univ. Grenoble, from Jan 2013 until Dec 2015]
Philippe Weinzaepfel [Univ. Grenoble, from Nov 2012 until August 2016]

**Post-Doctoral Fellows**

Anoop Cherian [Inria, from Sep 2012 until Sep 2014]
Albert Gordo [Inria, MBDA France, from Aug 2012 until Jan 2014]
Piotr Koniusz [Inria, from Jul 2013 until March 2015]
Heng Wang [Inria, from July 2012 until April 2014]

**Visiting Scientists**

Adam Bloniarz [UC Berkeley, from Jul 2013 until Aug 2013]
Georgia Gkioxari [UC Berkeley, from Jul 2013 until Sep 2013]
Alexander Goldenshluger [ENSIMAG, from Mar 2013 until Jun 2013]
Thanh Tam Le [Kyoto University, Mar 2013]
Miles Lopes [UC Berkeley, from Apr 2013 until Jul 2013]
Jitendra Malik [UC Berkeley, from Jul 2013 until Aug 2013]
Hyun Oh Song [UC Berkeley, from Aug 2013 until Nov 2013]

**Administrative Assistant**

Nathalie Gillot [Inria]

**Others**

Yuansi Chen [Inria, intern, from Apr 2013 until Aug 2013]

Théo Trouillon [Inria, intern, from Feb 2013 until Jun 2013]
Michael Guerzhoy [Inria, intern, until Aug 2013]

# 2. Overall Objectives

## 2.1. Introduction

LEAR's main focus is learning-based approaches to visual object recognition and scene interpretation, particularly for object category detection, image retrieval, video indexing and the analysis of humans and their movements. Understanding the content of everyday images and videos is one of the fundamental challenges of computer vision, and we believe that significant advances will be made over the next few years by combining state-of-the-art image analysis tools with emerging machine learning and statistical modeling techniques.

LEAR's main research areas are:

- **Robust image descriptors and large-scale search.** Many efficient lighting and viewpoint invariant image descriptors are now available, such as affine-invariant interest points and histogram of oriented gradient appearance descriptors. Our research aims at extending these techniques to obtain better characterizations of visual object classes, for example based on 3D object category representations, and at defining more powerful measures for visual salience, similarity, correspondence and spatial relations. Furthermore, to search in large image datasets we aim at developing efficient correspondence and search algorithms.
- **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.
- **Visual category recognition.** Visual category recognition requires the construction of exploitable visual models of particular objects and of categories. Achieving good invariance to viewpoint, lighting, occlusion and background is challenging even for exactly known rigid objects, and these difficulties are compounded when reliable generalization across object categories is needed. Our research combines advanced image descriptors with learning to provide good invariance and generalization. Currently the selection and coupling of image descriptors and learning techniques is largely done by hand, and one significant challenge is the automation of this process, for example using automatic feature selection and statistically-based validation. Another option is to use complementary information, such as text, to improve the modeling and learning process.
- **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.

## 2.2. Highlights of the Year

- **TrecVid Multimedia Event Detection challenge.** We participated in the Multimedia Event Detection track of TrecVid 2013, one of the major benchmarks in automatic video analysis. We ranked first out of 18 participants [35].
- **ICCV'13 THUMOS Challenge.** We participated in the action recognition challenge THUMOS, organized in conjunction with ICCV '13. We were ranked first among 16 participants.
- **Optical Flow Benchmark SINTEL.** Our optical flow method DeepFlow [31] was ranked first to the online evaluation benchmark SINTEL from Max Planck Institute.
- **Cor Baayen Award.** Julien Mairal received the Cor Baayen prize, which is awarded annually by ERCIM to a promising young researcher in the field of Informatics and Applied Mathematics.
- **Best Phd prize.** Thomas Mensink, a former PhD student of LEAR, was awarded the best PhD thesis prize from AFRIF.

# 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). In a collaboration with Xerox Research Center Europe, supported by a CIFRE grant from ANRT, we study large-scale image annotation. In the context of the Microsoft-Inria collaboration we concentrate on retrieval and auto-annotation of videos by combining textual information (scripts accompanying videos) with video descriptors. In the EU FP7 project AXES we will further mature such video annotation techniques, and apply them to large archives in collaboration with partners such as the BBC, Deutsche Welle, and the Netherlands Institute for Sound and Vision.

**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. Our industrial project with MBDA is on detecting objects under such conditions in infrared images.

# 5. Software and Platforms

## 5.1. Large-scale image classification

**Participants:** Matthijs Douze [correspondant], Zaid Harchaoui, Florent Perronnin [XRCE], Cordelia Schmid.

JSGD is the implementation of a Stochastic Gradient Descent algorithm used to train linear multiclass classifiers. It is biased towards large classification problems (many classes, many examples, high-dimensional data). It can be used on the ImageNet large scale classification challenge. It uses several optimization techniques, both algorithmic (scale factors to spare vector multiplications, vector compression with product quantizers) and technical (vector operations, multithreading, improved cache locality). It has Python and Matlab interfaces. It is distributed under a Cecill licence. Project page: http://lear.inrialpes.fr/src/jsgd.

## 5.2. Fisher vector image representation

**Participants:** Matthijs Douze [correspondant], Hervé Jégou [TEXMEX Team Inria Rennes], Cordelia Schmid.

We developed a package that computes Fisher vectors on sparse or dense local SIFT features. The dense feature extraction was optimized, so that they can be computed in real time on video data. The implementation was used for several publications and in our submisssion to the Trecvid 2013 MED task. We provide a binary version of the local descriptor implementation, and the Fisher implementation is integrated in the Yael library, with Python and Matlab interface, see http://lear.inrialpes.fr/src/inria_fisher.

## 5.3. Video descriptors

**Participants:** Clement Leray, Dan Oneata, Cordelia Schmid [correspondant], Heng Wang, Jakob Verbeek.

We have developed and made on-line available software for video description based on dense trajectories and motion boundary histograms. The trajectories capture the local motion information of the video. A state-of-the-art optical flow algorithm enables a robust and efficient extraction of the dense trajectories. Descriptors are aligned with the trajectories and based on motion boundary histograms (MBH) which are robust to camera motion. This year we have further developed this software to increase its robustness and scalability to large datasets. Most importantly, we have added a robust background stabilization technique, which allows to remove camera motion. This has shown to significantly improve the performance. Furthermore, we have improved the efficiency of the approach. For example, we avoid writing the raw MBH descriptors to disk, but rather aggregate them directly into a signature for the complete video using Fisher vectors. This allowed us to use these descriptors on the 4,000 hour video dataset of the TrecVid 2013 MED task as well as on the 3500 hours of AXES broadcast videos.

## 5.4. SPArse Modeling Software (SPAMS)

**Participants:** Julien Mairal [correspondant], Jean-Paul Chieze [WILLOW Project-Team], Jean Ponce [WILLOW Project-Team], Francis Bach [SIERRA Project-Team].

SPAMS v2.4 was released as open-source software in December 2013 (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. A graphical tool for visualizing dictionaries was developed by Jean-Paul Chieze, and stochastic optimization tools corresponding to the papers [24], [23] were added for dealing with large-scale sparse estimation problems.

## 5.5. 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 [36]. 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 softwares 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 1.0.0 is a user friendly bioconductor R package. It is freely available on the Bioconductor website under a GPL licence: http://bioconductor.org/packages/release/bioc/html/flipflop.html.

## 5.6. DeepFlow

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

We developed a package for the "deep flow" algorithm [31]. "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.

## 5.7. Object category localization

**Participants:** Ramazan Cinbis, Matthijs Douze, Cordelia Schmid, Jakob Verbeek.

We developed an object category localization system based on a Fisher vector representation over densly extracted local SIFT descriptors [18]. To improve the robustness with respect to background clutter in the detection windows we developed an approximate object segmentation method that is used to weigh the contribution of local SIFT descriptors. Our system achieves state-of-the-art localization performance as measured on the PASCAL VOC 2007 and 2010 datasets. The system is developed in both C, python, and Matlab. The system will be released in early 2014.

# 6. New Results

## 6.1. Visual recognition in images

### 6.1.1. Label-Embedding for Attribute-Based Classification

**Participants:** Zeynep Akata, Florent Perronnin, Zaid Harchaoui, Cordelia Schmid.

Attributes are an intermediate representation, which enables parameter sharing between classes, a must when training data is scarce. We propose in [13] to view attribute-based image classification as a label-embedding problem: each class is embedded in the space of attribute vectors. We introduce a function which measures the compatibility between an image and a label embedding, as shown in Figure 1. The parameters of this function are learned on a training set of labeled samples to ensure that, given an image, the correct classes rank higher than the incorrect ones. Results on the Animals With Attributes and Caltech-UCSD-Birds datasets show that the proposed framework outperforms the standard Direct Attribute Prediction baseline in a zero-shot learning scenario. The label embedding framework offers other advantages such as the ability to leverage alternative sources of information in addition to attributes (e.g. class hierarchies) or to transition smoothly from zero-shot learning to learning with large quantities of data.

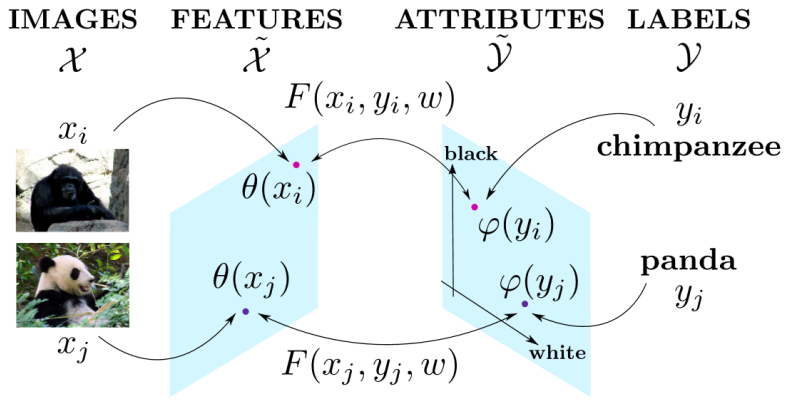



*Figure 1. Much work in computer vision has been devoted to image embedding (left): how to extract suitable features from an image? We focus on label embedding (right): how to embed class labels in a Euclidean space? We use attributes as side information for the label embedding and measure the "compatibility" between the embedded inputs and outputs with a function F*

### 6.1.2. Good Practice in Large-Scale Learning for Image Classification

**Participants:** Zeynep Akata, Florent Perronnin, Zaid Harchaoui, Cordelia Schmid.

In this paper [2], 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%.

### *6.1.3. Segmentation Driven Object Detection with Fisher Vectors*

**Participants:** Ramazan Gokberk Cinbis, Jakob Verbeek, Cordelia Schmid.

In [18], we present an object detection system based on the Fisher vector (FV) image representation computed over SIFT and color descriptors. For computational and storage efficiency, we use a recent segmentation-based method to generate class-independent object detection hypotheses, in combination with data compression techniques. Our main contribution is a method to produce tentative object segmentation masks to suppress background clutter in the features. As illustrated in Figure 2, re-weighting the local image features based on these masks is shown to improve object detection significantly. We also exploit contextual features in the form of a full-image FV descriptor, and an inter-category rescoring mechanism. Our experiments on the VOC 2007 and 2010 datasets show that our detector improves over the current state-of-the-art detection results.

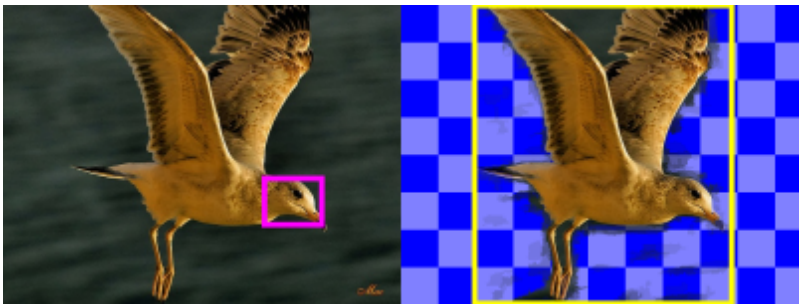


*Figure 2. The image on the left and the one on the right show the top detection without and with using our segmentation-driven descriptors, respectively.*

### *6.1.4. Image Classification with the Fisher Vector: Theory and Practice*

**Participants:** Jorge Sánchez, Florent Perronnin, Thomas Mensink, Jakob Verbeek.

A standard approach to describe an image for classification and retrieval purposes is to extract a set of local patch descriptors, encode them into a high-dimensional vector and pool them into an image-level signature. The most common patch encoding strategy consists in quantizing the local descriptors into a finite set of prototypical elements. This leads to the popular Bag-of-Visual words (BOV) representation. In [10], we propose to use the Fisher Kernel framework as an alternative patch encoding strategy: we describe patches by their deviation from a "universal" generative Gaussian mixture model. This representation, which we call Fisher Vector (FV) has many advantages: it is efficient to compute, it leads to excellent results even with efficient linear classifiers, and it can be compressed with a minimal loss of accuracy using product

quantization. We report experimental results on five standard datasets – PASCAL VOC 2007, Caltech 256, SUN 397, ILSVRC 2010 and ImageNet10K – with up to 9M images and 10K classes, showing that the FV framework is a state-of-the-art patch encoding technique. In figure 3 we show a representative benchmark performance comparison between BOV and FV representations.

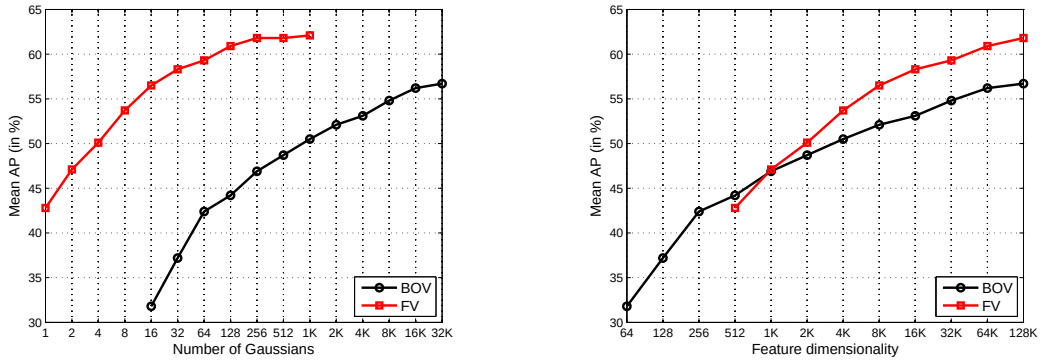



*Figure 3. Accuracy of the BOV and the FV as a function of the number of Gaussians (left) and feature dimensionality (right) on PASCAL VOC 2007 with SIFT descriptors only.*

## 6.2. Learning and statistical models

### *6.2.1. Kernel-Based Methods for Hypothesis Testing: A Unified View*

**Participants:** Zaid Harchaoui, Francis Bach, Olivier Cappe, Eric Moulines.

Kernel-based methods provide a rich and elegant framework for developing nonparametric detection procedures for signal processing. Several recently proposed procedures can be simply described using basic concepts of reproducing kernel Hilbert space embeddings of probability distributions, namely mean elements and covariance operators. In [5], we propose a unified view of these tools, and draw relationships with information divergences between distributions (see Figure 4).

### *6.2.2. Supervised Feature Selection in Graphs with Path Coding Penalties and Network Flows*

**Participants:** Julien Mairal, Bin Yu.

In this paper [6], we consider supervised learning problems where the features are embedded in a graph, such as gene expressions in a gene network. In this context, it is of much interest to automatically select a subgraph with few connected components; by exploiting prior knowledge, one can indeed improve the prediction performance or obtain results that are easier to interpret. Regularization or penalty functions for selecting features in graphs have recently been proposed, but they raise new algorithmic challenges. For example, they typically require solving a combinatorially hard selection problem among all connected subgraphs. In this paper, we propose computationally feasible strategies to select a sparse and well-connected subset of features sitting on a directed acyclic graph (DAG), see Figure 5. We introduce structured sparsity penalties over paths on a DAG called "path coding" penalties. Unlike existing regularization functions that model long-range interactions between features in a graph, path coding penalties are tractable. The penalties and their proximal operators involve path selection problems, which we efficiently solve by leveraging network flow optimization. We experimentally show on synthetic, image, and genomic data that our approach is scalable and leads to more connected subgraphs than other regularization functions for graphs.

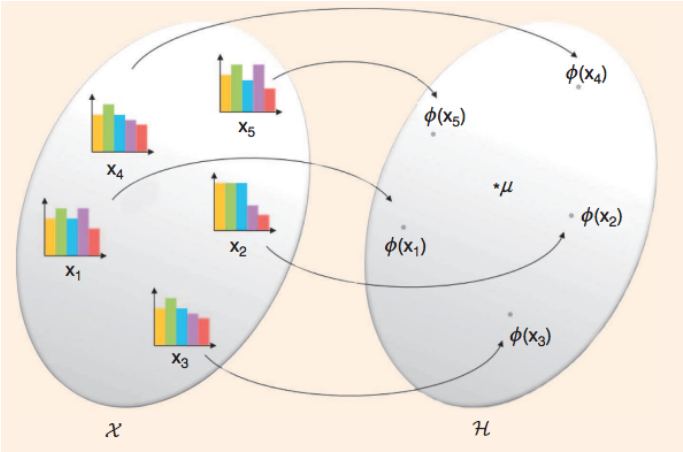



*Figure 4. A schematic view of kernel embedding and mean element*

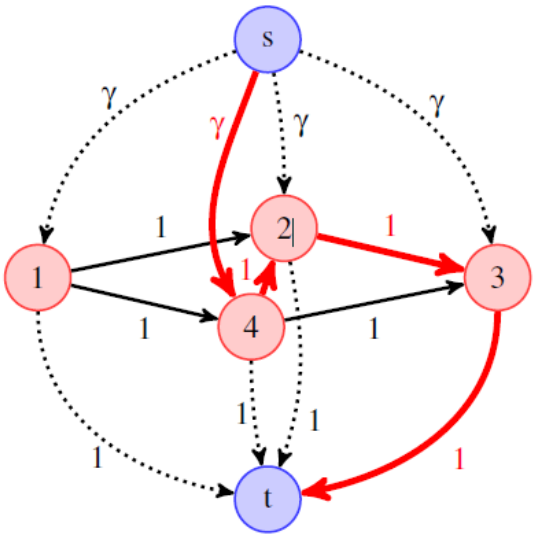
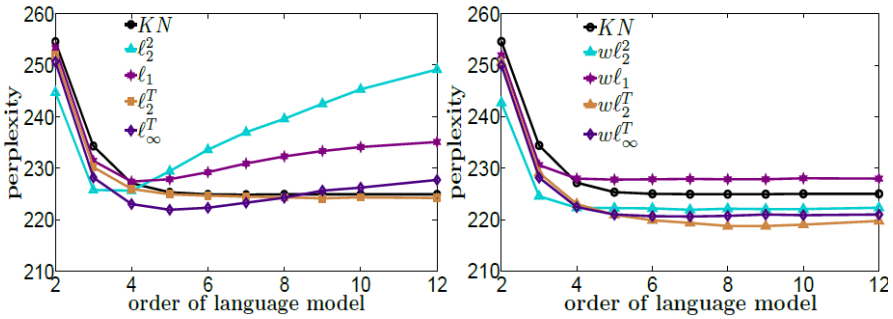



*Figure 5. Network Flow Model with Costs on Arcs for the Path Selection Problem*

### 6.2.3. Structured Penalties for Log-linear Language Models

**Participants:** Anil Nelakanti, Cédric Archambeau, Julien Mairal, Francis Bach, Guillaume Bouchard.

Language models can be formalized as loglinear regression models where the input features represent previously observed contexts up to a certain length $m$. The complexity of existing algorithms to learn the parameters by maximum likelihood scale linearly in $nd$, where $n$ is the length of the training corpus and $d$ is the number of observed features. In this paper [26], we present a model that grows logarithmically in $d$, making it possible to efficiently leverage longer contexts (see Figure 6). We account for the sequential structure of natural language using tree-structured penalized objectives to avoid overfitting and achieve better generalization.




(a) Unweighted penalties. (b) Weighted penalties.

*Figure 6. The classical measure of performance for natural language models is the perplexity (lower is better). Our models are denoted by $\ell_2^T$ and $\ell_{\text{inf}}^T$.*

### 6.2.4. Optimization with First-Order Surrogate Functions

**Participant:** Julien Mairal.

In this paper [23], we study optimization methods consisting of iteratively minimizing surrogates of an objective function, as illustrated in Figure 7. By proposing several algorithmic variants and simple convergence analyses, we make two main contributions. First, we provide a unified viewpoint for several first-order optimization techniques such as accelerated proximal gradient, block coordinate descent, or Frank-Wolfe algorithms. Second, 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.

### 6.2.5. Stochastic Majorization-Minimization Algorithms for Large-Scale Optimization

**Participant:** Julien Mairal.

Majorization-minimization algorithms consist of iteratively minimizing amajorizing surrogate of an objective function. Because of its simplicity and its wide applicability, this principle has been very popular in statistics and in signal processing. In this paper [24], we intend to make this principle scalable. We introduce a stochastic majorization-minimization scheme which is able to deal with largescale or possibly infinite data sets. When applied to convex optimization problems under suitable assumptions, we show that it achieves an expected convergence rate of $O(1/\sqrt{n})$ after $n$ iterations, and of $O(1/n)$ for strongly convex functions. Equally important, our scheme almost surely converges to stationary points for a large class of non-convex problems. We develop several efficient algorithms based on our framework. First, we propose a new stochastic proximal gradient method, which experimentally matches state-of-the-art solvers for large-scale $\ell_1$- logistic regression. Second, we develop an online DC programming algorithm for non-convex sparse estimation. Finally, we demonstrate the effectiveness of our approach for solving large-scale structured matrix factorization problems.

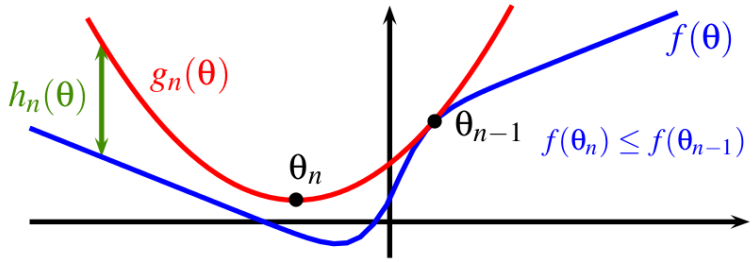



*Figure 7. Illustration of the basic majorization-minimization principle. We compute a surrogate $g_n$ of the objective function $f$ around a current estimate $\theta_{n-1}$. The new estimate $\theta_n$ is a minimizer of $g_n$. The approximation error $h_n$ is smooth.*

## 6.3. Recognition in video

### *6.3.1. Temporal Localization of Actions with Actoms*

**Participants:** Adrien Gaidon, Zaid Harchaoui, Cordelia Schmid.

In this paper [4], we address the problem of localizing actions, such as opening a door, in hours of challenging video data. We propose a model based on a sequence of atomic action units, termed "actoms", that are semantically meaningful and characteristic for the action. Our Actom Sequence Model (ASM) represents an action as a sequence of histograms of actom-anchored visual features, which can be seen as a temporally structured extension of the bag-of-features. Training requires the annotation of actoms for action examples. At test time, actoms are localized automatically based on a non-parametric model of the distribution of actoms, which also acts as a prior on an action's temporal structure. We present experimental results on two recent benchmarks for action localization "Coffee and Cigarettes" and the "DLSBP" dataset. We also adapt our approach to a classification-by-localization set-up, and demonstrate its applicability on the challenging "Hollywood 2" dataset. We show that our ASM method outperforms the current state of the art in temporal action localization, as well as baselines that localize actions with a sliding window method (see Figure 8).

### *6.3.2. Activity representation with motion hierarchies*

**Participants:** Adrien Gaidon, Zaid Harchaoui, Cordelia Schmid.

Complex activities, e.g., pole vaulting, are composed of a variable number of sub-events connected by complex spatio-temporal relations, whereas simple actions can be represented as sequences of short temporal parts. In [3], we learn hierarchical representations of activity videos in an unsupervised manner. These hierarchies of mid-level motion components are data-driven decompositions specific to each video. We introduce a spectral divisive clustering algorithm to efficiently extract a hierarchy over a large number of tracklets (i.e., local trajectories). We use this structure to represent a video as an unordered binary tree. We model this tree using nested histograms of local motion features. We provide an efficient positive definite kernel that computes the structural and visual similarity of two hierarchical decompositions by relying on models of their parent-child relations. We present experimental results on four recent challenging benchmarks: the High Five dataset, the Olympics Sports dataset, the Hollywood 2 dataset, and the HMDB dataset. We show that per-video hierarchies provide additional information for activity recognition. Our approach improves over unstructured activity models, baselines using other motion decomposition algorithms, and the state of the art (see Figure 9).

### *6.3.3. DeepFlow: Large displacement optical flow with deep matching*

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

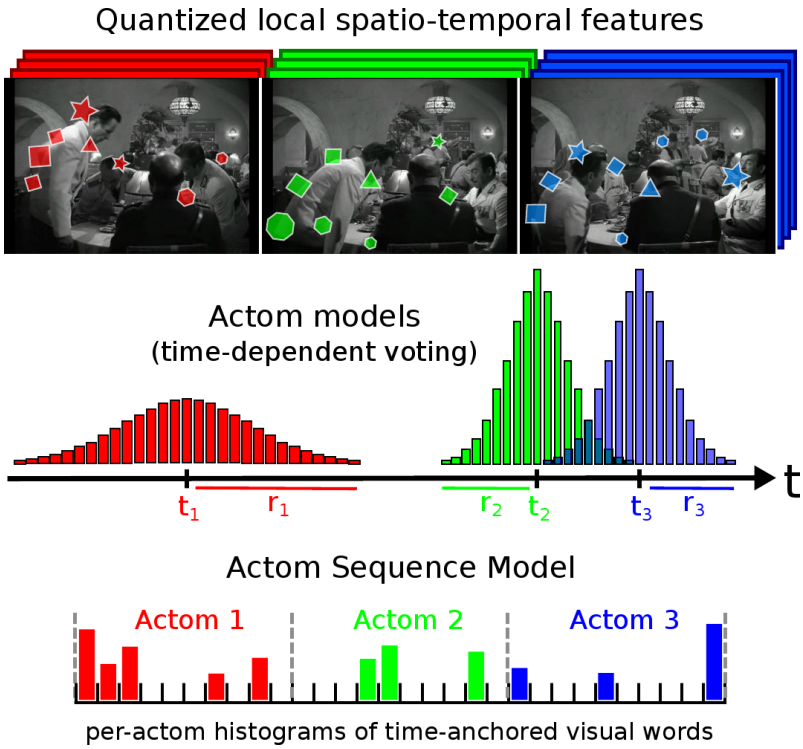



*Figure 8. Illustration of actoms-based decomposition of actions.*

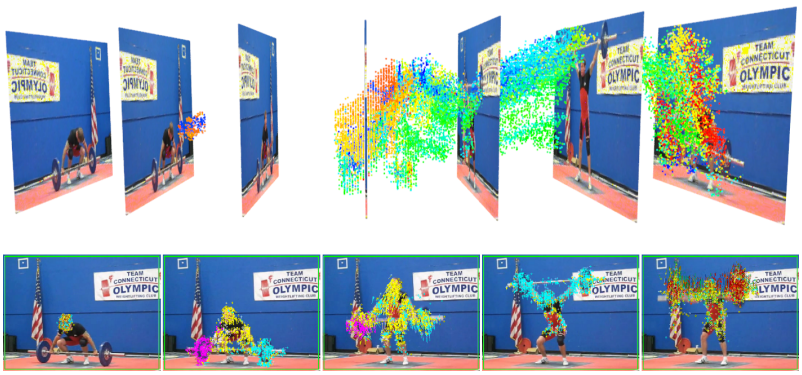


*Figure 9. Illustration of motion hierarchies for weight-lifting.*

Optical flow computation is a key component in many computer vision systems designed for tasks such as action detection or activity recognition. However, despite several major advances over the last decade, handling large displacement in optical flow remains an open problem. Inspired by the large displacement optical flow of Brox and Malik, our approach, termed DeepFlow, blends a matching algorithm with a variational approach for optical flow. We propose in [31] a descriptor matching algorithm, tailored to the optical flow problem, that allows to boost performance on fast motions. The matching algorithm builds upon a multi-stage architecture with 6 layers, interleaving convolutions and max-pooling, a construction akin to deep convolutional nets. Figure 10 shows an outline of our approach. Using dense sampling, it allows to efficiently retrieve quasi-dense correspondences, and enjoys a built-in smoothing effect on descriptors matches, a valuable asset for integration into an energy minimization framework for optical flow estimation. DeepFlow efficiently handles large displacements occurring in realistic videos, and shows competitive performance on optical flow benchmarks. Furthermore, it sets a new state-of-the-art on the MPI-Sintel dataset.

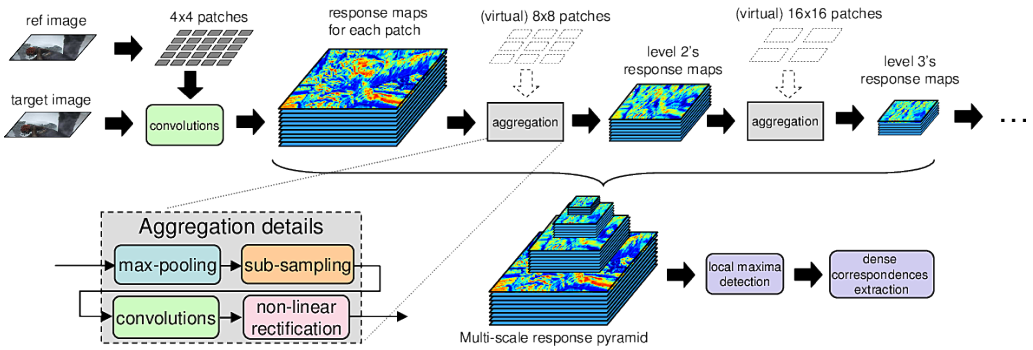



*Figure 10. Outline of DeepFlow.*

### *6.3.4. Event retrieval in large video collections with circulant temporal encoding*

**Participants:** Jerome Revaud, Matthijs Douze, Cordelia Schmid, Hervé Jégou.

This paper [28] presents an approach for large-scale event retrieval. Given a video clip of a specific event, e.g., the wedding of Prince William and Kate Middleton, the goal is to retrieve other videos representing the same event from a dataset of over 100k videos. Our approach encodes the frame descriptors of a video to jointly represent their appearance and temporal order. It exploits the properties of circulant matrices to compare the videos in the frequency domain. This offers a significant gain in complexity and accurately localizes the matching parts of videos, see Figure 11. Furthermore, we extend product quantization to complex vectors in order to compress our descriptors, and to compare them in the compressed domain. Our method outperforms the state of the art both in search quality and query time on two large-scale video benchmarks for copy detection, Trecvid and CCweb. Finally, we introduce a challenging dataset for event retrieval, EVVE, and report the performance on this dataset.

### *6.3.5. Dense trajectories and motion boundary descriptors for action recognition*

**Participants:** Heng Wang, Alexander Kläser, Cordelia Schmid, Cheng-Lin Liu.

This paper [11] introduces a video representation based on dense trajectories and motion boundary descriptors. Trajectories capture the local motion information of the video. A state-of-the-art optical flow algorithm enables a robust and efficient extraction of the dense trajectories. As descriptors we extract features aligned with the trajectories to characterize shape (point coordinates), appearance (histograms of oriented gradients) and motion (histograms of optical flow). Additionally, we introduce a descriptor based on motion boundary

*Figure 11. Example of correctly aligned videos. Each row is a different video, and each column corresponds to temporally aligned frames from the videos.*

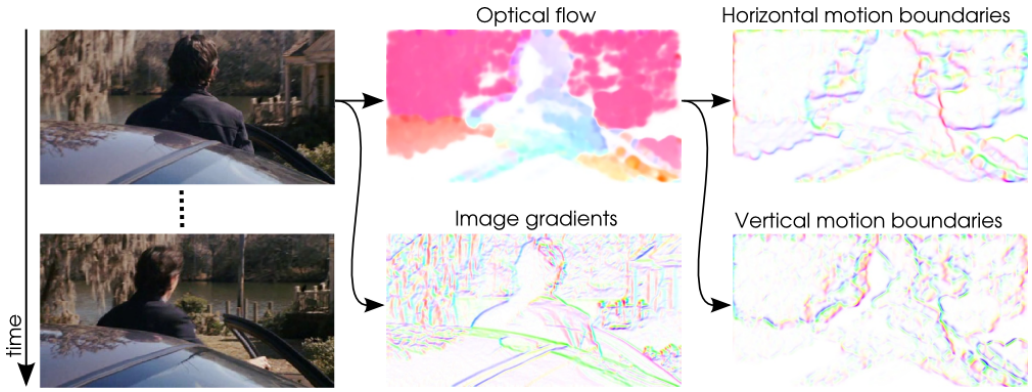



*Figure 12. Illustration of the information captured by HOG, HOF, and MBH descriptors. Gradient/flow orientation is indicated by color (hue) and magnitude by saturation. The optical flow (top, middle) shows constant motion in the background, which is due to the camera movements. The motion boundaries (right) encode the relative motion between the person and the background.*

histograms (MBH) (see the visualization in Figure 12), which is shown to consistently outperform other state-of-the-art descriptors, in particular on real-world videos that contain a significant amount of camera motion. We evaluate our video representation in the context of action classification on nine datasets, namely KTH, YouTube,Hollywood2, UCF sports, IXMAS, UIUC, Olympic Sports, UCF50 and HMDB51. On all datasets our approach outperforms current state-of-the-art results.

### *6.3.6. Action Recognition with Improved Trajectories*

**Participants:** Heng Wang, Cordelia Schmid.

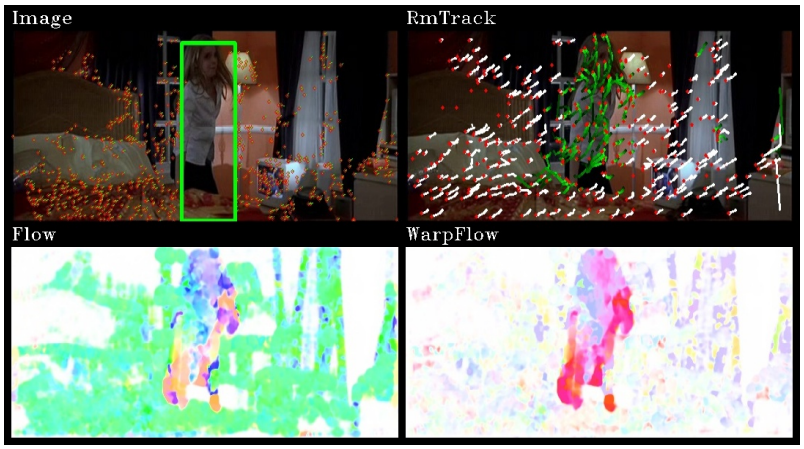



*Figure 13. Visualization of human detection and inlier matches (top, left) as well as removed background trajectories, which are due to camera motion (top, right). The bottom row compares the original optical flow (bottom, left) and the warped version (bottom, right).*

This paper [30] improves dense trajectories by taking into account camera motion to correct them. To estimate camera motion, we match feature points between frames using SURF descriptors and dense optical flow, which are shown to be complementary. These matches are, then, used to robustly estimate a homography with RANSAC. Human motion is in general different from camera motion and generates inconsistent matches. To improve the estimation, a human detector is employed to remove these matches. Given the estimated camera motion, we remove trajectories consistent with it. We also use this estimation to cancel out camera motion from the optical flow. This significantly improves motion-based descriptors, such as HOF and MBH (see Figure 13). Experimental results on four challenging action datasets (i.e., Hollywood2, HMDB51, Olympic Sports and UCF50) significantly outperform the current state of the art.

### *6.3.7. Action and event recognition with Fisher vectors on a compact feature set*

**Participants:** Dan Oneaţă, Jakob Verbeek, Cordelia Schmid.

Action recognition in uncontrolled video is an important and challenging computer vision problem. Recent progress in this area is due to new local features and models that capture spatio-temporal structure between local features, or human-object interactions. Instead of working towards more complex models, we focus in this paper [27] on the low-level features and their encoding. We evaluate the use of Fisher vectors as an alternative to bag-of-word histograms to aggregate a small set of state-of-the-art low-level descriptors, in combination with linear classifiers. We present a large and varied set of evaluations, considering (i) classification of short actions in five datasets, (ii) localization of such actions in feature-length movies, and (iii) large-scale recognition of complex events. We find that for basic action recognition and localization MBH features alone are enough for state-of-the-art performance. For complex events we find that SIFT and MFCC

features provide complementary cues. On all three problems we obtain state-of-the-art results, while using fewer features and less complex models.

### 6.3.8. *Stable hyper-pooling and query expansion for event detection*

**Participants:** Matthijs Douze, Jerome Revaud, Cordelia Schmid, Hervé Jégou.

This work [19] makes two complementary contributions to event retrieval in large collections of videos. First, we compare different ways of quantizing video frame descriptors in terms of temporal stability. Our best choices compare favorably with the standard pooling technique based on k-means quantization, see Figure 14. Second, we introduce a technique to improve the ranking. It can be interpreted either as a query expansion method or as a similarity adaptation based on the local context of the query video descriptor. Experiments on public benchmarks show that our methods are complementary and improve event retrieval results, without sacrificing efficiency.

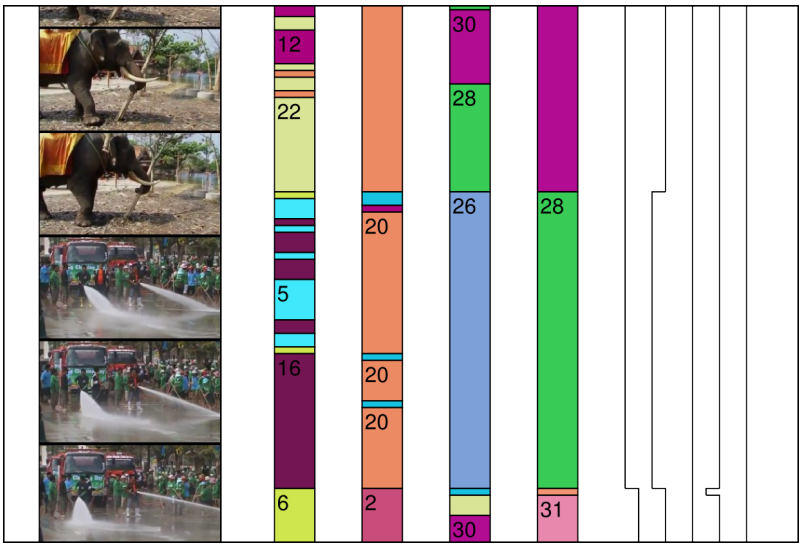



*Figure 14. Several quantizations of video frame descriptors (left) to a color-coded index in* $\{0, ..., 31\}$. *Leftmost column: standard k-means, right: the proposed SSC. Time runs vertically.*

### 6.3.9. *Finding Actors and Actions in Movies.*

**Participants:** Piotr Bojanowski, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, Josef Sivic.

This work [16] addresses the problem of learning a joint model of actors and actions in movies using weak supervision provided by scripts. Specifically, we extract actor/action pairs from the script and use them as constraints in a discriminative clustering framework. The corresponding optimization problem is formulated as a quadratic program under linear constraints. People in video are represented by automatically extracted and tracked faces together with corresponding motion features. First, we apply the proposed framework to the task of learning names of characters in movies and demonstrate significant improvements over previous methods used for this task. Second, we explore joint actor/action constraints and show their advantage for weakly supervised action learning. We validate our method in the challenging setting of localizing and recognizing characters and their actions in the feature length movies Casablanca and American Beauty. Figure 15 shows an example of our results.

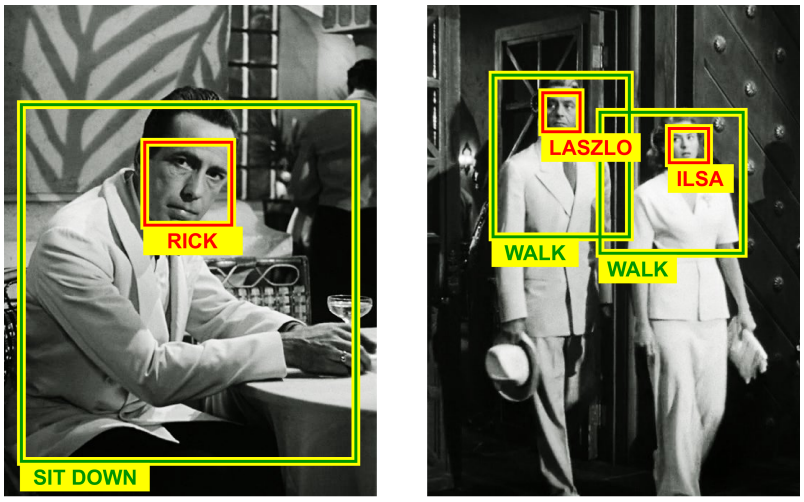



*Figure 15. Automatic detection and annotation of characters and their actions in the movie Casablanca. The automatically resolved correspondence between video and script is color-coded.*

# 7. Bilateral Contracts and Grants with Industry

## 7.1. MBDA Aerospatiale

**Participants:** Albert Gordo, Michael Guerzhoy, Cordelia Schmid, Franck Thollard.

The collaboration with the Aerospatiale section of MBDA has been on-going for several years: MBDA has funded the PhD of Yves Dufurnaud (1999-2001), a study summarizing the state-of-the-art on recognition (2004), a one year transfer contract on matching and tracking (11/2005-11/2006) as well as the PhD of Hedi Harzallah (2007-2010). From September 2010 to 2013, we conducted a three-year contract on object localization and pose estimation based on shape representation.

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

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

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.

## 7.3. 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.

## 7.4. Xerox Research Center Europe

**Participants:** Zeynep Akata, Zaid Harchaoui, Cordelia Schmid.

The collaboration with Xerox started in October 2009 with a co-supervised CIFRE scholarship (2009-2012) on cross-modal information retrieval. A second three-year collaborative project on large scale visual recognition started in 2011. The goal is to design algorithms for large-scale image classification possibly in the presence of missing labels. The joint PhD student Zeynep Akata is supported by a CIFRE grant obtained from the ANRT. She graduated in early January 2014.

# 8. Partnerships and Cooperations

## 8.1. National Initiatives

### *8.1.1. QUAERO Project*

**Participants:** Matthijs Douze, Dan Oneata, Danila Potapov, Jerome Revaud, Cordelia Schmid, Franck Thollard, Heng Wang.

Quaero is a French-German search engine project supported by OSEO. It runs from 2008 to 2013 and includes many academic and industrial partners, such as Inria, CNRS, the universities of Karlsruhe and Aachen as well as LTU, Exalead and INA. LEAR/Inria is involved in the tasks of automatic image annotation, image clustering as well as large-scale image and video search. See http://www.quaero.org for more details.

### *8.1.2. 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.

### *8.1.3. 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.

### *8.1.4. 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].

Several inverse problems in genomics involve retrieving meaningful DNA sequences from observed data. This is for example the case of the isoform deconvolution problem of RNA-Seq data, which is currently of utmost importance in genomics. The problem can be cast as a sparse feature selection problem, where the features are mapped to the paths of a graph called "splicing graph". Even though the number of paths is exponential in the graph size, we investigate network flow optimization techniques to efficiently solve the inverse problem in polynomial time [36]. The project involves researchers in machine learning, optimization, bio-informatics, and biology, from Inria Rhone-Alpes, Institut Curie in Paris, and the LBBE laboratory in Lyon.

### *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. 2013

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

**Participant:** Zaid Harchaoui.

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 will start in Jan. 2014 and end in Dec. 2016.

### *8.1.7. Project Math-STIC "Gauge"*

**Participant:** Zaid Harchaoui.

The project is concerned with statistical learning with gauge regularization penalty, a project funded by the Math-STIC "pôle" of the Université Joseph Fourier (Grenoble University). The partners are Inria Rhone-Alpes, CREST-ENSAE, Université Paris-Est. Principal investigator/leader: Zaid Harchaoui
Dates: Jan 2012-Dec 2013.

## 8.2. European Initiatives

### *8.2.1. FP7 European Project AXES*

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

This 4-year project started in January 2011. 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. FP7 European Network of Excellence PASCAL 2*

**Participants:** Zeynep Akata, Adrien Gaidon, Zaid Harchaoui, Cordelia Schmid, Jakob Verbeek.

PASCAL (Pattern Analysis, Statistical Modeling and Computational Learning) is a 7th framework EU Network of Excellence that started in March 2008 for five years. It has established a distributed institute that brings together researchers and students across Europe, and is now reaching out to countries all over the world. PASCAL is developing the expertise and scientific results that will help create new technologies such as intelligent interfaces and adaptive cognitive systems. To achieve this, it supports and encourages collaboration between experts in machine learning, statistics and optimization. It also promotes the use of machine learning in many relevant application domains such as machine vision. The project ended in February 2013.

### *8.2.3. ERC Advanced grant Allegro*

**Participants:** Cordelia Schmid, Karteek Alahari, Jerome Revaud.

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.** Inria principal investigator: Zaid Harchaoui. International Partner (Institution - Laboratory - Researcher): University of California Berkeley (United States) - Electrical Engineering and Computer Science Department. Duration: 2012 - 2014. The goal of the associated team "Hyperion" is to take up the challenges of large-scale statistical learning for image interpretation and video understanding. Despite the ever-increasing number of large annotated image and video datasets, designing principled and scalable statistical learning approaches from such big computer vision datasets remains a major scientific challenge.

  The associated team consists of researchers from the LEAR project team of Inria and two teams of University of California Berkeley (resp. the Pr. Jitendra Malik and the Pr. Nourredine El Karoui teams). It allows the three teams to effectively combine their respective strengths in areas such as large-scale learning theory and algorithms, high-level feature design for computer vision, and high-dimensional statistical learning theory. It will result in significant progress in domains such as large-scale image classification, weakly-supervised learning for classification into attributes, and transfer learning.

### *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. Adam Bloniarz, who is pursuing his PhD under the supervision of Prof. Bin Yu, visited LEAR duing the summer 2013.
- **University of Edinburgh:** C. Schmid collaborates with V. Ferrari, associate professor at university of Edinburgh. Our initial collaboration (co-supervision of A. Prest 2009-2012) was renewed this year. 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. This resulted in two ICCV'13 publications: one on modeling pose with flexible human puppets [32] and one on measuring the impact of low, intermediate and high-level descriptions on action recognition [22]. C. Schmid plans to continue this collaboration in 2014.

### *8.3.3. Participation In other International Programs*

- **France-Berkeley fund:** The LEAR team was awarded in 2012 a grant from the France-Berkeley fund for the project with Pr. Jitendra Malik (EECS, UC Berkeley) on "Large-scale learning for image and video interpretation". The award amounts to 10,000 USD for a period of one year, from September 2012 to September 2013. 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 of International Scientists*

- Jitendra Malik, Professor in UC Berkeley, visited LEAR during the summer 2013 as part of the associated team "Hyperion" and a project from the France-Berkeley fund. The goal of his visit was to develop new approaches for human action classification and localization in videos.

### *8.4.2. Internships*

- Georgia Gkioxari, a PhD student from UC Berkeley, visited LEAR during the summer 2013 as part of the associated team "Hyperion" and a project from the France-Berkeley fund. The goal of her visit was to develop new approaches for human action localization in videos.
- Hyun Oh Song, a PhD student from UC Berkeley, visited LEAR during the fall 2013 as part of the associated team "Hyperion". The goal of his visit was to develop efficient approaches for part-based models in computer vision.
- Miles Lopes, a PhD student from UC Berkeley, visited LEAR during the spring 2013 as part of the associated team "Hyperion". The goal of his visit was to develop efficient approaches for estimating statistical functionals using convex optimization.
- Adam Bloniarz, a PhD student from UC Berkeley, visited LEAR during the summer 2013 as part of the associated team "Hyperion". The goal of his visit was to develop video representations adapted to neuroscience, based upon computer vision principles.

# 9. Dissemination

## 9.1. Scientific Animation

- Conference, workshop, and summer school organization:
  - Z. Harchaoui has co-organized the workshop “Optimization and Statistical Learning” at Les Houches, January 2013.
  - Z. Harchaoui and F. Pierucci: Co-organizer of the NIPS 2013 workshop on Greedy Algorithms, Frank-Wolfe and Friends - A modern perspective. December 2013.
  - C. Schmid and Z. Harchaoui: Co-organizers of the Inria Visual Recognition and Machine Learning Summer School, Paris, July 2013. Attracted a total of 177 participants from 34 countries (49 participants from France, 57 from Europe and 71 from America, Asia, and Africa).
- Editor-in-chief:
  - C. Schmid: International Journal of Computer Vision, since 2013.
- Editorial boards:
  - K. Alahari: Guest editor for the Special Issue on “Higher Order Graphical Models in Computer Vision: Modelling, Inference and Learning” to be published in IEEE Transactions on Pattern Analysis and Machine Intelligence.
  - J. Mairal: Guest editor for the Special Issue on Sparse Coding to be published in the International Journal of Computer Vision.
  - C. Schmid: Foundations and Trends in Computer Graphics and Vision, since 2005.
  - J. Verbeek: Image and Vision Computing Journal, since 2011.
- General chair:
  - C. Schmid: CVPR ’15.
- Area chair:

  - C. Schmid: CVPR 2013, ICCV 2013.
  - J. Verbeek: BMVC 2013.
- Program committees (reviewers):
  - CVPR 2013: J. Verbeek, J. Mairal, K. Alahari, H. Wang, A. Gordo, Z. Harchaoui.
  - ICCV 2013: J. Verbeek, K. Alahari, H. Wang, A. Gordo.
  - NIPS 2013: J. Verbeek, J. Mairal, K. Alahari, Z. Harchaoui.
  - ICML 2013: J. Mairal, Z. Harchaoui.
  - IJCAI 2013: J. Verbeek.
  - BMVC 2013: K. Alahari, A. Gordo.
  - AISTATS 2013: Z. Harchaoui.
- Prizes:
  - Z. Harchaoui was awarded a Fellowship from the Isaac Newton Institute for Mathematical Sciences (Cambridge University) to participate to the "Inference for Change-Point and Related Processes" program in January 2014.
  - J. Mairal received the Cor Baayen prize, which is awarded annually by ERCIM to a promising young researcher in the field of Informatics and Applied Mathematics.
  - T. Mensink (former PhD student of LEAR) received the best PhD prize from AFRIF.

## 9.2. Teaching - Supervision - Juries

### *9.2.1. Teaching*

Doctorat: Z. Harchaoui co-organized a tutorial on "Large-Scale Visual Recognition" at CVPR, June 2013.

Doctorat: Z. Harchaoui, Tutorial on large-scale learning, 1h, Inria Visual Recognition and Machine Learning Summer School 2013, Paris, France.

Doctorat: J. Mairal, Tutorial for the IMA Short Course "Applied Statistics and Machine Learning", 3H, Minneapolis, June 2013.

Doctorat: C. Schmid, Tutorial on image search and classification, 3h, Inria Visual Recognition and Machine Learning Summer School 2013, Paris, France.

Master: M. Douze, K. Alahari, Bases de donnees multimedia, M2, ENSIMAG, France..

Master: Z. Harchaoui, Kernel-based methods for statistical machine learning, 18H, M2, Université Joseph Fourier (Grenoble University);

Master: J. Mairal and Z. Harchaoui. Statistical Learning and Applications. 16H, M2, Ecole Normale Supérieure de Lyon, France.

Master: C. Schmid, Object recognition and computer vision, 10h, M2, ENS ULM, France.

Master: C. Schmid and J. Verbeek, Machine Learning & Category Representation, 18h, M2, ENSIMAG, France.

Master: P. Weinzaepfel, "Réseaux IP", 18H TD, M1, University Joseph Fourier.

Licence: F. Pierucci, "Calculus", 75H TD, L1, University Joseph Fourier.

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

Licence: P. Weinzaepfel, "Communications numériques", 12H TD, L1, Polytech Grenoble.

### *9.2.2. Supervision*

PhD: Z. Akata, Contributions to Large-Scale Learning for Image Classification, Université de Grenoble, 06/01/2014, advisors: C. Schmid and F. Perronnin.

### *9.2.3. Juries*

- J. Mairal, reviewer and jury member for the PhD thesis of Roberto Rigamonti, EPFL, Lausanne, December 18th, 2013.
- C. Schmid, présidente du jury concours CR2 Inria Grenoble, May 2013.
- C. Schmid, jury member of PhD committee for Nicolas Ballas, Ecole Nationale Supérieure des Mines de Paris, November 2013.
- C. Schmid, jury member of PhD committee for Wafa Bel Haj Ali, Université Nice-Sophia Antipolis, October 2013.
- C. Schmid, jury member of HDR committee for Ivan Laptev, Ecole Normale Superieure, July 2013.
- C. Schmid, jury member of PhD committee for Sandra Avila, Université Pierre et Marie Curie, June 2013.
- J. Verbeek, reviewer and jury member of PhD committee for K. Simonyan, University of Oxford, November 6 2013.

## 9.3. Invited presentations

- K. Alahari: Invited speaker at the Brookes Anniversary Worshop, University of Oxford, Oxford, UK, October 2013.
- K. Alahari: Invited speaker at the Maori Workshop, Ecole Polytechnique, Palaiseau, France, November 2013:
- K. Alahari: Seminar at IIIT Hyderabad, India, December 2013.
- Z. Harchaoui: Seminar at GDR Isis, Paris, October 2013.
- Z. Harchaoui: talk at ICCOPT 2013, Portugal, July 2013.
- Z. Harchaoui: Seminar at CREST-ENSAE, Paris, June 2013.
- Z. Harchaoui: invited speaker at SMAI Conference, Seignosse, May 2013.
- Z. Harchaoui: invited speaker at Optimization and big data workshop, Edinburgh, May 2013.
- Z. Harchaoui: invited speaker at IHES, Orsay, March 2013.
- J. Mairal: Invited speaker at the Maori Workshop, Ecole Polytechnique, Palaiseau, France, November 2013.
- J. Mairal: Cor Baayen award presentation at ERCIM meeting, Athens, October 2013.
- J. Mairal: Seminar at Gipsa-Lab, Grenoble, June 2013.
- J. Mairal: Seminar FREMIT, Toulouse, April 2013.
- J. Mairal: Seminar at Xerox Research Center Europe (XRCE), Grenoble, March 2013.
- F. Perucci: talk at ICCOPT 2013, Portugal, July 2013.
- C. Schmid: Seminar at MPI, Tubingen, January 2013.
- C. Schmid: CVPR area chair meeting workshop, USC, February 2013.
- C. Schmid: Seminar at UCF, Orlando, May 2013.
- C. Schmid: Seminar at WILLOW retreat, Bandol, June 2013.
- C. Schmid: Invited speaker at Second international workshop on visual analysis and geo-localizaton of large-scale imagery in conjunction with CVPR'13, June 2013.
- C. Schmid: Keynote speaker at 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS), Paris, July 2013.
- C. Schmid: ICCV area chair meeting workshop, Oxford University, August 2013.
- C. Schmid: Keynote talk at The First International Workshop on Action Recognition with a Large Number of Classes, in conjunction with ICCV '13, Sydney, Australia, December 2013.

- J. Verbeek: Invited speaker at DGA workshop on Multimedia Information Processing (TIM 2013), Paris, France, July 2013.
- J. Verbeek: Seminar at Media Integration and Communication Center, University of Florence, Italy, September 2013.
- J. Verbeek: Seminar at Intelligent Systems Lab Amsterdam, University of Amstedam, The Netherlands, October 2013.

## 9.4. Popularization

- C. Schmid presented the research area of visual recognition to a group of high-school teachers, Inria Grenoble, March 2013.
- A. Gaidon, Z. Harchaoui and C. Schmid published an article about activity recognition in videos in ERCIM-news 95 (October 2013).

# 10. Bibliography

## Publications of the year

### Doctoral Dissertations and Habilitation Theses

[1] Z. AKATA. , *Contributions à l'apprentissage grande échelle pour la classification d'images*, Université de Grenoble, January 2014, http://hal.inria.fr/tel-00873807

### Articles in International Peer-Reviewed Journals

[2] 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", June 2013 [*DOI :* 10.1109/TPAMI.2013.146], http://hal.inria.fr/hal-00835810

[3] A. GAIDON, Z. HARCHAOUI, C. SCHMID. *Activity representation with motion hierarchies*, in "International Journal of Computer Vision", November 2013 [*DOI :* 10.1007/s11263-013-0677-1], http://hal.inria.fr/hal-00908581

[4] A. GAIDON, Z. HARCHAOUI, C. SCHMID. *Temporal Localization of Actions with Actoms*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", November 2013, vol. 35, n$^{o}$ 11, pp. 2782-2795 [*DOI :* 10.1109/TPAMI.2013.65], http://hal.inria.fr/hal-00804627

[5] Z. HARCHAOUI, F. BACH, O. CAPPÉ, E. MOULINES. *Kernel-Based Methods for Hypothesis Testing: A Unified View*, in "IEEE Signal Processing Magazine", June 2013, vol. 30, n$^{o}$ 4, pp. 87-97 [*DOI :* 10.1109/MSP.2013.2253631], http://hal.inria.fr/hal-00841978

[6] J. MAIRAL, B. YU. *Supervised Feature Selection in Graphs with Path Coding Penalties and Network Flows*, in "Journal of Machine Learning Research", August 2013, vol. 14, pp. 2449-2485, http://hal.inria.fr/hal-00806372

[7] T. MENSINK, J. VERBEEK, G. CSURKA. *Tree-structured CRF Models for Interactive Image Labeling*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", February 2013, vol. 35, n$^{o}$ 2, pp. 476-489 [*DOI :* 10.1109/TPAMI.2012.100], http://hal.inria.fr/hal-00688143

[8] T. MENSINK, J. VERBEEK, F. PERRONNIN, G. CSURKA. *Distance-Based Image Classification: Generalizing to new classes at near-zero cost*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", November 2013, vol. 35, n$^{o}$ 11, pp. 2624-2637 [*DOI :* 10.1109/TPAMI.2013.83], http://hal.inria.fr/hal-00817211

[9] A. PREST, V. FERRARI, C. SCHMID. *Explicit modeling of human-object interactions in realistic videos*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", April 2013, vol. 35, n$^{o}$ 4, pp. 835-848 [*DOI :* 10.1109/TPAMI.2012.175], http://hal.inria.fr/hal-00720847

[10] J. SANCHEZ, F. PERRONNIN, T. MENSINK, J. VERBEEK. *Image Classification with the Fisher Vector: Theory and Practice*, in "International Journal of Computer Vision", December 2013, vol. 105, n$^{o}$ 3, pp. 222-245 [*DOI :* 10.1007/s11263-013-0636-x], http://hal.inria.fr/hal-00830491

[11] H. WANG, A. KLÄSER, C. SCHMID, C.-L. LIU. *Dense trajectories and motion boundary descriptors for action recognition*, in "International Journal of Computer Vision", May 2013, vol. 103, n$^{o}$ 1, pp. 60-79 [*DOI :* 10.1007/s11263-012-0594-8], http://hal.inria.fr/hal-00803241

### International Conferences with Proceedings

[12] Z. AKATA, F. PERRONNIN, Z. HARCHAOUI, C. SCHMID. *Attribute-Based Classification with Label-Embedding*, in "NIPS 2013 Workshop on Output Representation Learning", Lake Tahoe, United States, Neural Information Processing Systems (NIPS) Foundation, December 2013, http://hal.inria.fr/hal-00903502

[13] Z. AKATA, F. PERRONNIN, Z. HARCHAOUI, C. SCHMID. *Label-Embedding for Attribute-Based Classification*, in "CVPR 2013 - IEEE Computer Vision and Pattern Recognition", Portland, United States, IEEE, June 2013, http://hal.inria.fr/hal-00815747

[14] K. ALAHARI, G. SEGUIN, J. SIVIC, I. LAPTEV. *Pose Estimation and Segmentation of People in 3D Movies*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, 2013, http://hal.inria.fr/hal-00874884

[15] J. ALMAZAN, A. GORDO, A. FORNÉS, E. VALVENY. *Handwritten Word Spotting with Corrected Attributes*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, December 2013, http://hal.inria.fr/hal-00906787

[16] P. BOJANOWSKI, F. BACH, I. LAPTEV, J. PONCE, C. SCHMID, J. SIVIC. *Finding Actors and Actions in Movies*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, 2013, http://hal.inria.fr/hal-00904991

[17] M. CHO, K. ALAHARI, J. PONCE. *Learning Graphs to Match*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, 2013, http://hal.inria.fr/hal-00875105

[18] R. G. CINBIS, J. VERBEEK, C. SCHMID. *Segmentation Driven Object Detection with Fisher Vectors*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, December 2013, http://hal.inria.fr/hal-00873134

[19] M. DOUZE, J. REVAUD, C. SCHMID, H. JÉGOU. *Stable hyper-pooling and query expansion for event detection*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, October 2013, http://hal.inria.fr/hal-00872751

[20] A. GAMAL ELDIN, G. CHARPIAT, X. DESCOMBES, J. ZERUBIA. *An efficient optimizer for simple point process models*, in "SPIE, Computational Imaging XI", Burlingame, California, United States, C. A. BOUMAN, I. POLLAK, P. J. WOLFE (editors), SPIE Proceedings, SPIE, February 2013, vol. 8657 [*DOI :* 10.1117/12.2009238], http://hal.inria.fr/hal-00801448

[21] A. GANDHI, K. ALAHARI, C. V. JAWAHAR. *Decomposing Bag of Words Histograms*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, 2013, http://hal.inria.fr/hal-00874895

[22] H. JHUANG, J. GALL, S. ZUFFI, C. SCHMID, M. J. BLACK. *Towards understanding action recognition*, in "ICCV 2013 - IEEE Intenational Conference on Computer Vision", Sydney, Australia, IEEE, December 2013, http://hal.inria.fr/hal-00906902

[23] J. MAIRAL. *Optimization with First-Order Surrogate Functions*, in "ICML 2013 - International Conference on Machine Learning", Atlanta, United States, JMLR Proceedings, June 2013, vol. 28, pp. 783-791, http://hal.inria.fr/hal-00822229

[24] J. MAIRAL. *Stochastic Majorization-Minimization Algorithms for Large-Scale Optimization*, in "NIPS - Advances in Neural Information Processing Systems", South Lake Tahoe, United States, December 2013, vol. 27, http://hal.inria.fr/hal-00835840

[25] A. MISHRA, K. ALAHARI, C. V. JAWAHAR. *Image Retrieval using Textual Cues*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, 2013, http://hal.inria.fr/hal-00875100

[26] A. NELAKANTI, C. ARCHAMBEAU, J. MAIRAL, F. BACH, G. BOUCHARD. *Structured Penalties for Log-linear Language Models*, in "EMNLP - Empirical Methods in Natural Language Processing - 2013", Seattle, United States, Association for Computational Linguistics, October 2013, pp. 233-243, http://hal.inria.fr/hal-00904820

[27] D. ONEATA, J. VERBEEK, C. SCHMID. *Action and Event Recognition with Fisher Vectors on a Compact Feature Set*, in "ICCV 2013 - IEEE Intenational Conference on Computer Vision", Sydney, Australia, IEEE, December 2013, http://hal.inria.fr/hal-00873662

[28] J. REVAUD, M. DOUZE, C. SCHMID, H. JÉGOU. *Event retrieval in large video collections with circulant temporal encoding*, in "CVPR 2013 - International Conference on Computer Vision and Pattern Recognition", Portland, United States, IEEE, March 2013, http://hal.inria.fr/hal-00801714

[29] G. SHARMA, F. JURIE, C. SCHMID. *Expanded Parts Model for Human Attribute and Action Recognition in Still Images*, in "CVPR 2013 - International Conference on Computer Vision and Pattern Recognition", Portland, Oregon, United States, IEEE, April 2013, pp. 652-659 [*DOI :* 10.1109/CVPR.2013.90], http://hal.inria.fr/hal-00816144

[30] H. WANG, C. SCHMID. *Action Recognition with Improved Trajectories*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, December 2013, http://hal.inria.fr/hal-00873267

[31] P. WEINZAEPFEL, J. REVAUD, Z. HARCHAOUI, C. SCHMID. *DeepFlow: Large displacement optical flow with deep matching*, in "ICCV 2013 - IEEE Intenational Conference on Computer Vision", Sydney, Australia, IEEE, December 2013, http://hal.inria.fr/hal-00873592

[32] S. ZUFFI, J. ROMERO, C. SCHMID, M. J. BLACK. *Estimating Human Pose with Flowing Puppets*, in "ICCV 2013 - IEEE Intenational Conference on Computer Vision", Sydney, Australia, IEEE, December 2013, http://hal.inria.fr/hal-00906800

### Research Reports

[33] P. KONIUSZ, F. YAN, P.-H. GOSSELIN, K. MIKOLAJCZYK. , *Higher-order Occurrence Pooling on Mid- and Low-level Features: Visual Concept Detection*, September 2013, 20 p. , http://hal.inria.fr/hal-00922524

[34] J. SANCHEZ, F. PERRONNIN, T. MENSINK, J. VERBEEK. , *Image Classification with the Fisher Vector: Theory and Practice*, Inria, May 2013, n$^{O}$ RR-8209, http://hal.inria.fr/hal-00779493

### Other Publications

[35] R. ALY, R. ARANDJELOVIC, K. CHATFIELD, M. DOUZE, B. FERNANDO, Z. HARCHAOUI, K. MCGUINESS, N. O'CONNOR, D. ONEATA, O. PARKHI, D. POTAPOV, J. REVAUD, C. SCHMID, J.-L. SCHWENNINGER, D. SCOTT, T. TUYTELAARS, J. VERBEEK, H. WANG, A. ZISSERMAN. , *The AXES submissions at TrecVid 2013*, November 2013, TRECVID Workshop, Gaithersburg, United States, http://hal.inria.fr/hal-00904404

[36] E. BERNARD, L. JACOB, J. MAIRAL, J.-P. VERT. , *Efficient RNA Isoform Identification and Quantification from RNA-Seq Data with Network Flows*, September 2013, http://hal.inria.fr/hal-00803134

[37] F. ENIKEEVA, Z. HARCHAOUI. , *High-dimensional change-point detection with sparse alternatives*, 2013, http://hal.inria.fr/hal-00933185