IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université Joseph Fourier (Grenoble)

Activity Report 2015

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

Creation of the Project-Team: 2003 July 01, end of the Project-Team: 2015 December 31

Keywords:

Computer Science and Digital Science:

3.4. - Machine learning and statistics
5.3. - Image processing and analysis
5.4. - Computer vision

Other Research Topics and Application Domains:

9.4.5. - Data science

1. Members

Research Scientists

Cordelia Schmid [Team leader, Inria, Senior Researcher, HdR]
Karteek Alahari [Inria, Researcher]
Zaid Harchaoui [Inria, Researcher, on leave at NYU since Oct. 2015]
Julien Mairal [Inria, Researcher, "en détachement du Corps des Mines"]
Jakob Verbeek [Inria, Researcher]

Engineers

Julien Bardonnet [Inria, funded by MBDA, from May 2015 to April 2016]
Matthijs Douze [Inria, SED 40 %, on leave at Facebook since Nov 2015]
Xavier Martin [Inria, funded by FP7 AXES project then by ERC Allegro, from Oct 2014 to Sep 2016]
Jerome Revaud [Inria, until Nov 2015, funded by ERC Allegro]

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 ERC Allegro, from Jul 2014 until Sep 2017]
Thomas Dias-Alves [Univ. Grenoble I, co-supervised with M. Blum (TIMC laboratory), from Oct 2014 to Sep 2017]
Yang Hua [Univ. Grenoble, funded by MSR/Inria joint lab, from Jan 2013 until June 2016]
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 March 2016]
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]
Konstantin Shmelkov [Univ. Grenoble, funded by ERC Allegro, from Oct 2015 until Oct 2018]
Pavel Tokmakov [Univ. Grenoble, funded by ERC Allegro, from Sep 2014 until Sep 2017]
Philippe Weinzaepfel [Univ. Grenoble, funded by Université Joseph Fourier, from Nov 2012 until Sep 2016]
Daan Wynen [Univ. Grenoble, funded by ERC Allegro and ANR Macaron, from Oct 2015 to Sep 2018]

Post-Doctoral Fellows

Anoop Cherian [Inria, funded by FP7 AXES project, from Nov 2012 until Feb 2015]
Piotr Koniusz [Inria, Inria Fellowship, from Jul 2013 until Mar 2015]
Guosheng Hu [Inria, funded by ANR PHYSIONOMIE project, from May 2015 to May 2016]
Yuri Maximov [INP Grenoble, until Jun 2015]

Marco Pedersoli [Inria, funded by ERC Allegro and MBDA project, from Sep 2015 until Sep 2016]
Xiaojiang Peng [Inria, funded by ERC Allegro, from Mar 2015 to Mar 2016]
Grégory Rogez [Inria, funded by FP7 Marie Curie IOF - Egovision4health, from Jul 2015 to Jun 2016]

Visiting Scientist
Yuansi Chen [UC Berkeley, from May to July 2015]

Administrative Assistant
Nathalie Gillot [Inria]

Others
Jérôme Lesaint [Inria, Intern, from Mar 2015 until Jun 2015]
Vladyslav Sydorov [Inria, Intern, funded by ERC Allegro, from Feb 2015 until March 2016]

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

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

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.

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. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

- Cordelia Schmid received the Humbolt research award, Alexander von Humbolt Foundation, Germany, 2015, and gave the Karen Spärck Jones lecture, annual event of the British Computer Society that honours women in computing research, 2015.
- Cordelia Schmid was ranked among the Thomson Reuters Highly Cited Researcher, 2015.
- Yang Hua, Karteek Alahari and Cordelia Schmid won the VOT-TIR2015 challenge.
- G. Cinbis (PhD, 2014) was awarded the 2014 AFRIF thesis prize for his thesis entitled "Fisher kernel based models for image classification and object localization" at Orasis 2015. He was supervised by Jakob Verbeek and Cordelia Schmid.
- N. Dalal (PhD, 2006) together with his supervisor B. Triggs was awarded the Longuet-Higgins Prize 2015 for his PhD work, in particular the paper entitled "Histograms of Oriented Gradients for Human Detection" (CVPR 2005 paper).

6. New Software and Platforms

6.1. Video descriptors

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

We have developed and made on-line available software for video description based on dense trajectories and motion boundary histograms, which are presented in [9]. 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. The code is available at http://lear.inrialpes.fr/~wang/improved_trajectories.

6.2. Patch CKN

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

This is an open-source software package implementing the image retrieval technique of [17]. It is available at http://lear.inrialpes.fr/people/paulin/projects/RomePatches/. The code relies on the software "Convolutional Kernel Networks" below.

6.3. Convolutional Kernel Networks

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

This is an open-source software package corresponding to a paper published at NIPS in 2014, available at http://ckn.gforge.inria.fr/, and which is continuously updated. 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.

6.4. 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", presented in the publication [31]. The code for "deep matching" is in python and the code for "deep flow" in C. The code is available on-line at http://lear.inrialpes.fr/src/deepmatching. In 2015, we have released a GPU version of "deep matching".

6.5. EpicFlow

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

We developed a package for the EpicFlow method [18], [32]. 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.

6.6. Motion Boundaries Detection

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

We make our source code for detecting motion boundaries [23] publicly available. The method is based on structured random forest and leverages both appearance and motion cues at the patch level. The source code is written in Matlab with C++ Mex-file and is available at http://lear.inrialpes.fr/research/motionboundaries/

6.7. Pose estimation and segmentation of multiple people

Participants: Guillaume Seguin, Karteek Alahari, Josef Sivic, Ivan Laptev.

We developed a method to obtain a pixel-wise segmentation and pose estimation of multiple people in stereoscopic videos. The codebase is composed of a set of patches for the various components in our pipeline, as well as the full pose mask generation and segmentation. It is available for download on the project website: http://www.di.ens.fr/willow/research/stereoseg.

6.8. 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. 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 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. In 2015, we released a new version to process multiple samples [4].

7. New Results

7.1. Visual recognition in images

7.1.1. Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning

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 [26], we propose to 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. 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. The technical report [26] is a journal paper under review after minor revision which extends a previous conference publication 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.

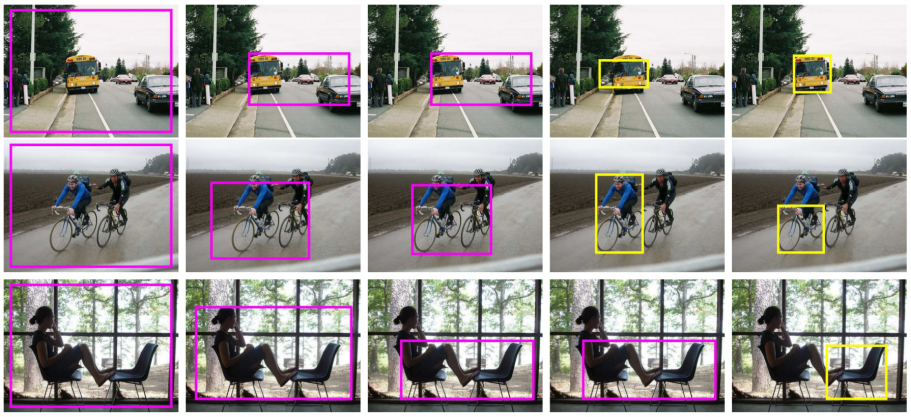

Figure 1. Illustration of our iterative object localization process on several example images, from initialization (left) to final localization (right). Yellow bounding boxes indicate that the object location hypothesis is in agreement with the ground-truth, for pink boxes the hypothesis is incorrect.

7.1.2. Patch-level spatial layout for classification and weakly supervised localization

Participants: Valentina Zadrija [University of Zagreb], Josip Krapac [University of Zagreb], Jakob Verbeek, Sinisa Segvic [University of Zagreb].

In [24] we propose a discriminative patch-level spatial layout model suitable for learning object localization models with weak supervision. We start from a block-sparse model of patch appearance based on the normalized Fisher vector representation. The appearance model is responsible for i) selecting a discriminative subset of visual words, and ii) identifying distinctive patches assigned to the selected subset. These patches are further filtered by a sparse spatial model operating on a novel representation of pairwise patch layout. We have evaluated the proposed pipeline in image classification and weakly supervised localization experiments on a public traffic sign dataset. The results show significant advantage of the proposed spatial model over state of the art appearance models.

7.1.3. Approximate Fisher Kernels of non-iid Image Models for Image Categorization

Participants: Ramazan Cinbis, Cordelia Schmid, Jakob Verbeek.

The bag-of-words (BoW) model treats images as sets of local descriptors and represents them by visual word histograms. The Fisher vector (FV) representation extends BoW, by considering the first and second order statistics of local descriptors. In both representations local descriptors are assumed to be identically and independently distributed (iid), which is a poor assumption from a modeling perspective. It has been experimentally observed that the performance of BoW and FV representations can be improved by employing discounting transformations such as power normalization. In [5], an expanded version of a previous conference publication, we introduce non-iid models by treating the model parameters as latent variables which are integrated out, rendering all local regions dependent. Using the Fisher kernel principle we encode an image by the gradient of the data log-likelihood w.r.t. the model hyper-parameters. Our models naturally generate discounting effects in the representations; suggesting that such transformations have proven successful because they closely correspond to the representations obtained for non-iid models. To enable tractable computation, we rely on variational free-energy bounds to learn the hyper-parameters and to compute approximate Fisher kernels. Our experimental evaluation results validate that our models lead to performance improvements comparable to using power normalization, as employed in state-of-the-art feature aggregation methods.

7.1.4. Local Convolutional Features with Unsupervised Training for Image Retrieval

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

Patch-level descriptors underlie several important computer vision tasks, such as stereo-matching or content-based image retrieval. We introduce a deep convolutional architecture that yields patch-level descriptors, as an alternative to the popular SIFT descriptor for image retrieval. The proposed family of descriptors, called Patch-CKN[17], adapt the recently introduced Convolutional Kernel Network (CKN), an unsupervised framework to learn convolutional architectures. We present a comparison framework to benchmark current deep convolutional approaches along with Patch-CKN for both patch and image retrieval (see Fig. 3 for our pipeline), including our novel "RomePatches" dataset. Patch-CKN descriptors yield competitive results compared to supervised CNNs alternatives on patch and image retrieval.

7.2. Learning and statistical models

7.2.1. A Universal Catalyst for First-order Optimization

Participants: Hongzhou Lin, Julien Mairal, Zaid Harchaoui.

In this paper [16], we introduce a generic scheme for accelerating first-order optimization methods in the sense of Nesterov, which builds upon a new analysis of the accelerated proximal point algorithm. Our approach consists of minimizing a convex objective by approximately solving a sequence of well-chosen auxiliary problems, leading to faster convergence. This strategy applies to a large class of algorithms, including gradient descent, block coordinate descent, SAG, SAGA, SDCA, SVRG, Finito/MISO, and their proximal variants. For all of these methods, we provide acceleration and explicit support for non-strongly convex objectives. In addition to theoretical speed-up, we also show that acceleration is useful in practice, as illustrated in Figure 4, especially for ill-conditioned problems where we measure significant improvements.

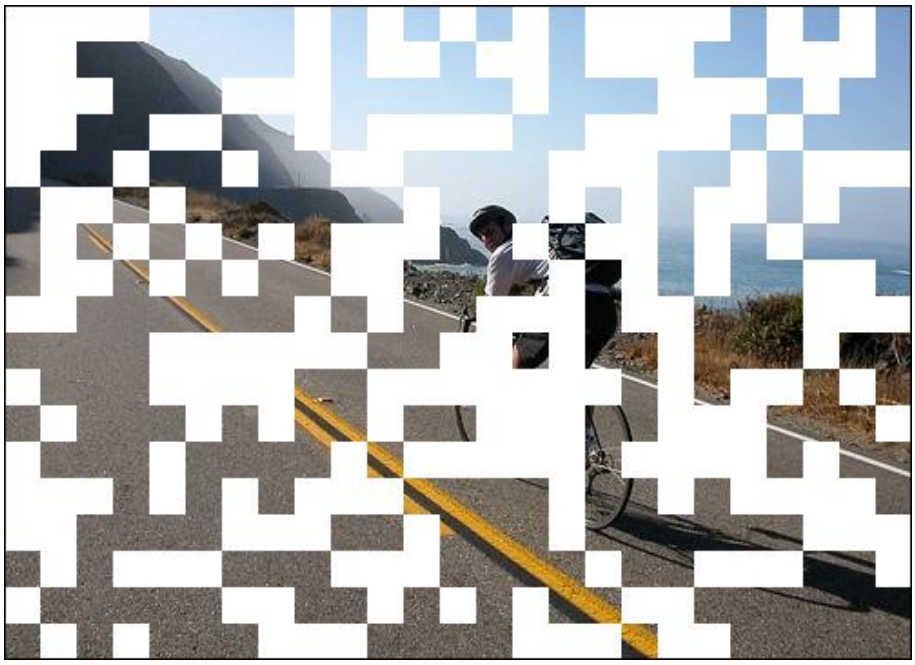

Figure 2. Illustration of why local image patches are not independent: we can easily guess the image content in the masked areas.

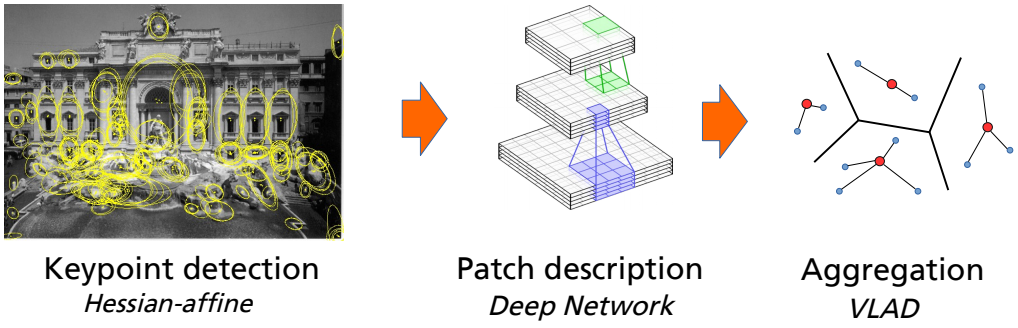

Figure 3. Image retrieval pipeline. Interest points are extracted with the Hessian-affine detector (left), encoded in descriptor space using convolutional features (middle), and aggregated into a compact representation using VLAD-pooling (right).

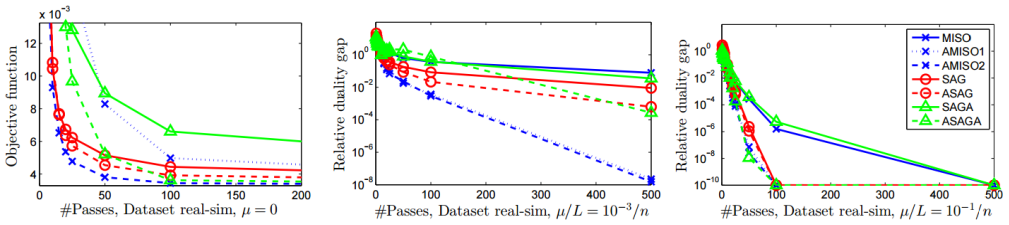

Figure 4. Objective function value (or duality gap) for different number of passes performed over each dataset. The legend for all curves is on the top right. AMISO, ASAGA, ASAG refer to the accelerated variants of MISO, SAGA, and SAG, respectively.

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

Participant: Julien Mairal.

In this paper [7], we study optimization methods consisting of iteratively minimizing surrogates of an objective function, as illustrated in Figure 5. 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.

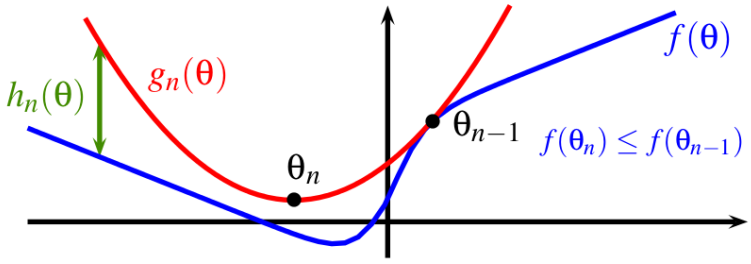

Figure 5. Illustration of the basic majorization-minimization principle. We compute a surrogate g_n of the objective function f around a current estimate θ_{n-1}. The new estimate θ_n is a minimizer of g_n. The approximation error h_n is smooth.

7.2.3. Coordinated Local Metric Learning

Participants: Shreyas Saxena, Jakob Verbeek.

Mahalanobis metric learning amounts to learning a linear data projection, after which the ℓ_2 metric is used to compute distances. In [20], we develop local metric learning techniques which allow more flexible metrics, not restricted to linear projections, see 6. Most of these methods partition the data space using clustering, and for each cluster a separate metric is learned. Using local metrics, however, it is not clear how to measure distances between data points assigned to different clusters. In this paper we propose to embed the local metrics in a global low-dimensional representation, in which the ℓ_2 metric can be used. With each cluster we associate a

linear mapping that projects the data to the global representation. This global representation directly allows computing distances between points regardless to which local cluster they belong. Moreover, it also enables data visualization in a single view, and the use of ℓ_2-based efficient retrieval methods. Experiments on the Labeled Faces in the Wild dataset show that our approach improves over previous global and local metric learning approaches.

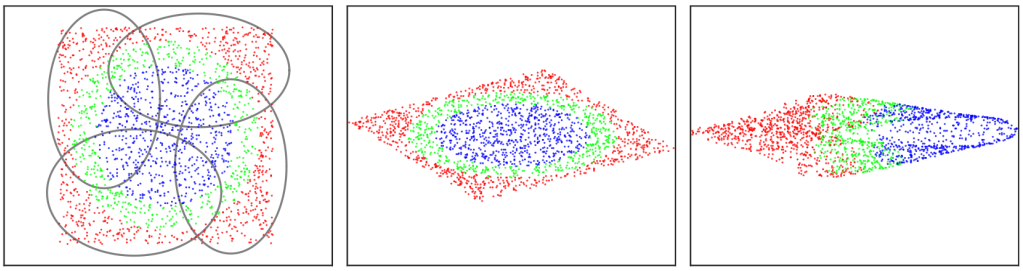

Figure 6. Synthetic dataset with color coded class labels, and the GMM used by our CLML local metric (left). Data projection given by a global Mahalanobis metric (middle) and our local CLML metric (right). The pairwise training constraints are better respected by CLML.

7.2.4. A convex formulation for joint RNA isoform detection and quantification from multiple RNA-seq samples

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].

Detecting and quantifying isoforms from RNA-seq data is an important but challenging task. The problem is often ill-posed, particularly at low coverage. One promising direction is to exploit several samples simultaneously. In this paper [4], we propose a new method for solving the isoform deconvolution problem jointly across several samples. We formulate a convex optimization problem that allows to share information between samples and that we solve efficiently, as illustrated in Figure 7. We demonstrate the benefits of combining several samples on simulated and real data, and show that our approach outperforms pooling strategies and methods based on integer programming. Our convex formulation to jointly detect and quantify isoforms from RNA-seq data of multiple related samples is a computationally efficient approach to leverage the hypotheses that some isoforms are likely to be present in several samples. The software and source code are available at http://cbio.ensmp.fr/flipflop.

7.2.5. Adaptive Recovery of Signals by Convex Optimization

Participants: Zaid Harchaoui, Anatoli Juditsky [Univ. Grenoble], Arkadi Nemirovski [Georgia Tech], Dimitry Ostrovsky [Univ. Grenoble].

In [13], we present a theoretical framework for adaptive estimation and prediction of signals of unknown structure in the presence of noise. The framework allows to address two intertwined challenges: (i) designing optimal statistical estimators; (ii) designing efficient numerical algorithms. In particular, we establish oracle inequalities for the performance of adaptive procedures, which rely upon convex optimization and thus can be efficiently implemented. As an application of the proposed approach, we consider denoising of harmonic oscillations

7.2.6. Semi-proximal Mirror-Prox for Nonsmooth Composite Minimization

Participants: Niao He [Georgia Tech], Zaid Harchaoui.

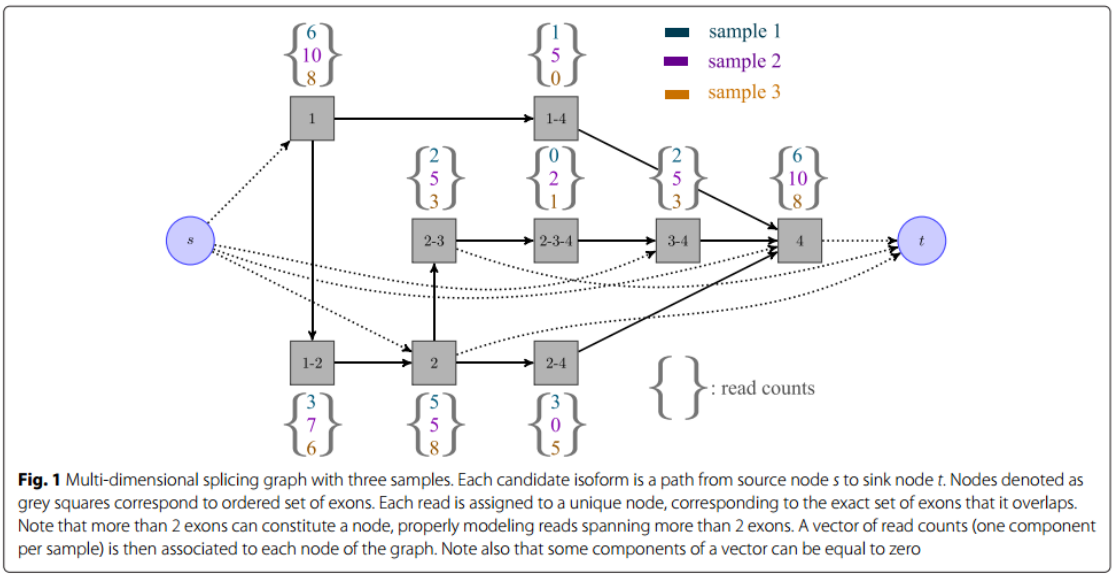

Fig. 1 Multi-dimensional splicing graph with three samples. Each candidate isoform is a path from source node *s* to sink node *t*. Nodes denoted as grey squares correspond to ordered set of exons. Each read is assigned to a unique node, corresponding to the exact set of exons that it overlaps. Note that more than 2 exons can constitute a node, properly modeling reads spanning more than 2 exons. A vector of read counts (one component per sample) is then associated to each node of the graph. Note also that some components of a vector can be equal to zero

Figure 7. Graph on which we perform network flow optimization. Nodes represent observed reads, and paths on the graph correspond to isoforms.

In [28], we propose a new first-order optimisation algorithm to solve high-dimensional non-smooth composite minimisation problems. Typical examples of such problems have an objective that decomposes into a non-smooth empirical risk part and a non-smooth regularisation penalty. The proposed algorithm, called Semi-Proximal Mirror-Prox, leverages the Fenchel-type representation of one part of the objective while handling the other part of the objective via linear minimization over the domain. The algorithm stands in contrast with more classical proximal gradient algorithms with smoothing, which require the computation of proximal operators at each iteration and can therefore be impractical for high-dimensional problems. We establish the theoretical convergence rate of Semi-Proximal Mirror-Prox, which exhibits the optimal complexity bounds, for the number of calls to linear minimization oracle. We present promising experimental results showing the interest of the approach in comparison to competing methods.

7.3. Recognition in video

7.3.1. Beat-Event Detection in Action Movie Franchises

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

While important advances were recently made towards temporally localizing and recognizing specific human actions or activities in videos, efficient detection and classification of long video chunks belonging to semantically-defined categories such as “pursuit” or “romance” remains challenging.

In our work [30], we introduce a new dataset, Action Movie Franchises, consisting of a collection of Hollywood action movie franchises. We define 11 non-exclusive semantic categories — called beat-categories — that are broad enough to cover most of the movie footage. The corresponding beat-events are annotated as groups of video shots, possibly overlapping. We propose an approach for localizing beat-events based on classifying shots into beat-categories and learning the temporal constraints between shots, as shown in Figure 8. We show that temporal constraints significantly improve the classification performance. We set up an evaluation protocol for beat-event localization as well as for shot classification, depending on whether movies from the same franchise are present or not in the training data.

Figure 8. A 5-minute extract from the proposed Action Movie Franchises dataset, ground truth annotation and output of different methods. Each color stands for a different event category: green —pursuit, blue —battle, yellow —victory-good, green —despair-good, pink —romance, gray —victory-bad, cadet blue —good-argue-good. Hashes mark difficult examples. The color code for the classifier evaluation is: white = true positive, gray = ignored, black = false positive.

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

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

In this paper [18], we propose a novel approach 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 – 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. An overview is given in Figure 9. EpicFlow significantly outperforms the state of the art on MPI-Sintel and performs on par on Kitti and Middlebury.

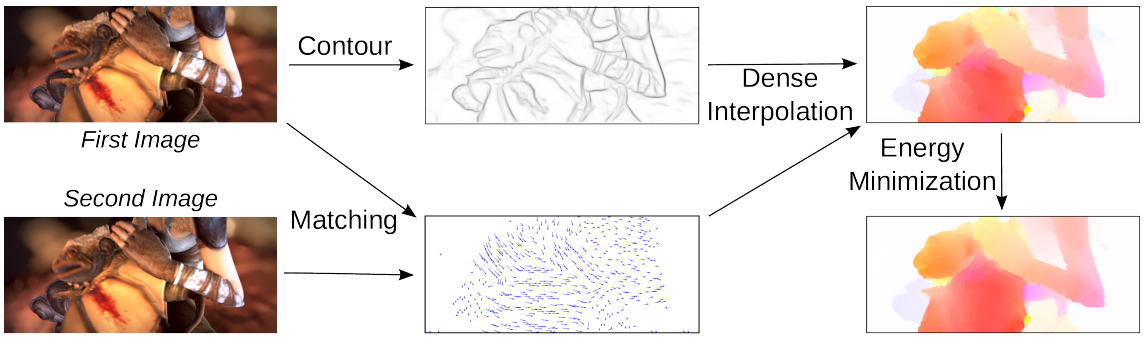

Figure 9. Overview of EpicFlow. Given two images, we compute matches using DeepMatching and the edges of the first image using SED. We combine these two cues to densely interpolate matches and obtain a dense correspondence field. This is used as initialization of a one-level energy minimization framework.

7.3.3. DeepMatching: Hierarchical Deformable Dense Matching

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

In this paper [31], we introduce a novel matching algorithm, called DeepMatching, to compute dense correspondences between images. DeepMatching relies on a hierarchical, multi-layer, correlational architecture designed for matching images and was inspired by deep convolutional approaches, see Figure 10. The proposed matching algorithm can handle non-rigid deformations and repetitive textures and efficiently determines dense correspondences in the presence of significant changes between images. We evaluate the performance of DeepMatching, in comparison with state-of-the-art matching algorithms, on the Mikolajczyk, the MPI-Sintel and the Kitti datasets. DeepMatching outperforms the state-of-the-art algorithms and shows excellent results in particular for repetitive textures. We also propose a method for estimating optical flow, called DeepFlow, by integrating DeepMatching in the large displacement optical flow (LDOF) approach of Brox et al. Compared to existing matching algorithms, additional robustness to large displacements and complex motion is obtained thanks to our matching approach. DeepFlow obtains competitive performance on public benchmarks for optical flow estimation.

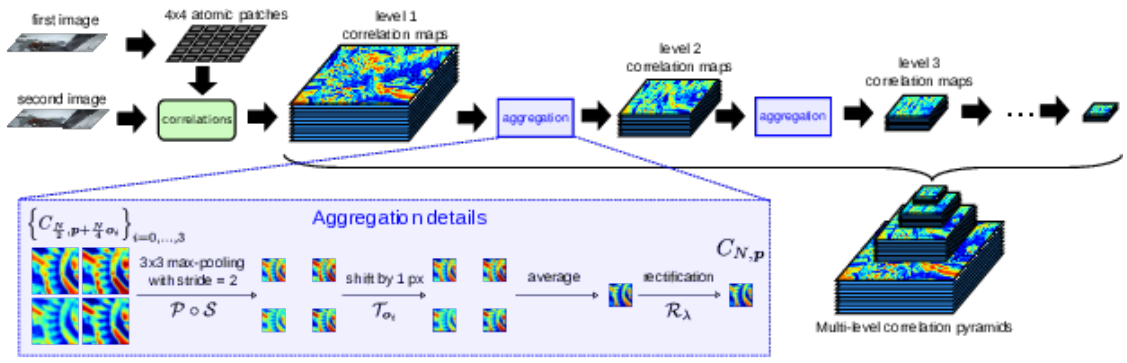

Figure 10. Overview of the bottom-up part of DeepMatching, which builds the multi-level correlation pyramid, from which matches are then extracted.

7.3.4. Learning to Detect Motion Boundaries

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

In this paper [23], we propose a learning-based approach for motion boundary detection. Precise localization of motion boundaries is essential for the success of optical flow estimation, as motion boundaries correspond to discontinuities of the optical flow field. The proposed approach allows to predict motion boundaries, using a structured random forest trained on the ground-truth of the MPI-Sintel dataset, see Figure 11. The random forest leverages several cues at the patch level, namely appearance (RGB color) and motion cues (optical flow estimated by state-of-the-art algorithms). Experimental results show that the proposed approach is both robust and computationally efficient. It significantly outperforms state-of-the-art motion-difference approaches on the MPI-Sintel and Middlebury datasets. We compare the results obtained with several state-of-the-art optical flow approaches and study the impact of the different cues used in the random forest. Furthermore, we introduce a new dataset, the YouTube Motion Boundaries dataset (**YMB**), that comprises 60 sequences taken from real-world videos with manually annotated motion boundaries. On this dataset, our approach, although trained on MPI-Sintel, also outperforms by a large margin state-of-the-art optical flow algorithms.

7.3.5. Learning to track for spatio-temporal action localization

Participants: Philippe Weinzaepfel, Zaid Harchaoui, Cordelia Schmid.

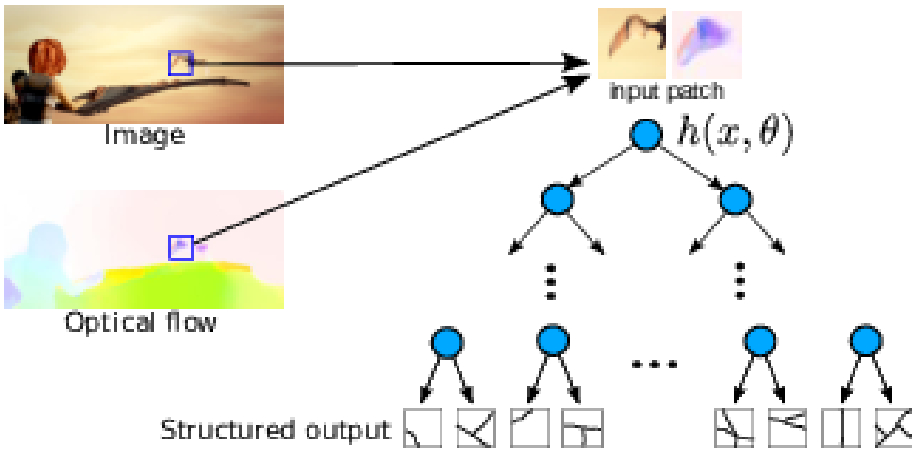

Figure 11. A structured random forest, taking as input a patch of an image and corresponding optical flow, outputs a motion boundaries patch. These motion boundaries patches are then aggregated to build the motion boundaries map for the whole image.

In this paper [22], we propose an effective approach for spatio-temporal action localization in realistic videos. The approach first detects proposals at the frame-level and scores them with a combination of static and motion CNN features. It then tracks high-scoring proposals throughout the video using a tracking-by-detection approach. Our tracker relies simultaneously on instance-level and class-level detectors. The tracks are scored using a spatio-temporal motion histogram, a descriptor at the track level, in combination with the CNN features. Finally, we perform temporal localization of the action using a sliding-window approach at the track level. An overview of our approach is given in Figure 12. We present experimental results for spatio-temporal localization on the UCF-Sports, J-HMDB and UCF-101 action localization datasets, where our approach outperforms the state of the art with a margin of 15%, 7% and 12% respectively in mAP.

7.3.6. A robust and efficient video representation for action recognition

Participants: Heng Wang, Dan Oneata, Cordelia Schmid, Jakob Verbeek.

In [9] we present a state-of-the-art video representation and apply it to efficient action recognition and detection. We first propose to improve the popular dense trajectory features by explicit camera motion estimation. Local feature trajectories consistent with the homography are considered as due to camera motion, and thus removed. This results in significant improvement on motion-based HOF and MBH descriptors. We further explore the recent Fisher vector as an alternative feature encoding approach to the standard bag-of-words histogram, and consider different ways to include spatial layout information in these encodings. We present a large and varied set of evaluations, considering (i) classification of short basic actions on six datasets, (ii) localization of such actions in featurelength movies, and (iii) large-scale recognition of complex events. We find that our improved trajectory features significantly outperform previous dense trajectories, and that Fisher vectors are superior to bag-of-words encodings for video recognition tasks. In all three tasks, we show substantial improvements over the state-of-the-art results. This journal paper combines and extends earlier conference papers.

7.3.7. Circulant temporal encoding for video retrieval and temporal alignment

Participants: Jerome Revaud, Matthijs Douze, Hervé Jégou [Inria Rennes, Facebook AI Research], Cordelia Schmid, Jakob Verbeek.

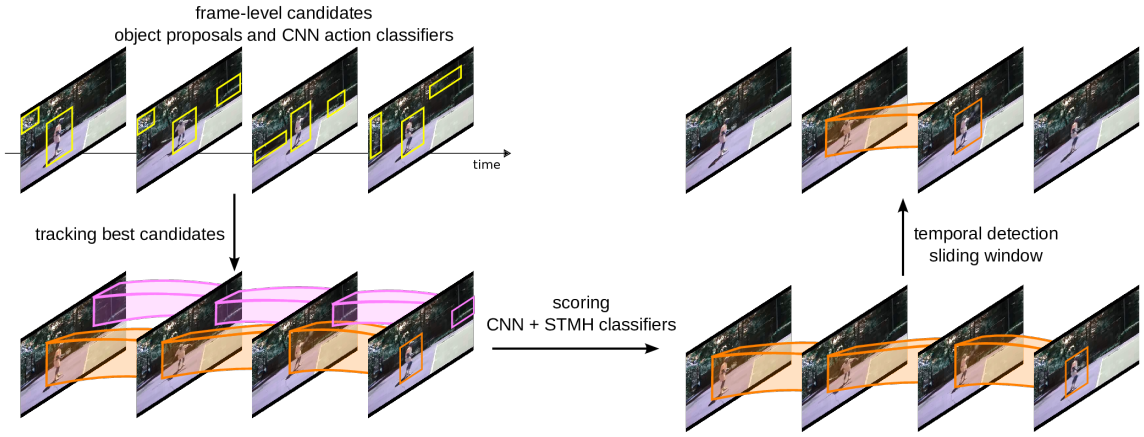

Figure 12. Overview of our spatio-temporal action localization approach. We detect frame-level object proposals and score them with CNN action classifiers. The best candidates, in term of scores, are tracked throughout the video. We then score the tracks with CNN and spatio-temporal motion histogram (STMH) classifiers. Finally, we perform a temporal sliding window for detecting the temporal extent of the action.

In [6] we address the problem of specific video event retrieval. Given a query video of a specific event, e.g., a concert of Madonna, the goal is to retrieve other videos of the same event that temporally overlap with the query. 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 efficiently compare the videos in the frequency domain. This offers a significant gain in complexity and accurately localizes the matching parts of videos. The descriptors can be compressed in the frequency domain with a product quantizer adapted to complex numbers. In this case, video retrieval is performed without decompressing the descriptors. The second problem we consider is the temporal alignment of a set of videos. We exploit the matching confidence and an estimate of the temporal offset computed for all pairs of videos by our retrieval approach. Our robust algorithm aligns the videos on a global timeline by maximizing the set of temporally consistent matches. The global temporal alignment enables synchronous playback of the videos of a given scene. This journal paper extends an earlier conference paper.

7.3.8. Pose Estimation and Segmentation of Multiple People in Stereoscopic Movies

Participants: Guillaume Seguin [Willow], Karteek Alahari, Josef Sivic [Willow], Ivan Laptev [Willow].

The work in [8] presents a method to obtain a pixel-wise segmentation and pose estimation of multiple people in stereoscopic videos, as shown in Figure 13. This task involves challenges such as dealing with unconstrained stereoscopic video, non-stationary cameras, and complex indoor and outdoor dynamic scenes with multiple people. We cast the problem as a discrete labelling task involving multiple person labels, devise a suitable cost function, and optimize it efficiently. The contributions of our work are two-fold: First, we develop a segmentation model incorporating person detections and learnt articulated pose segmentation masks, as well as colour, motion, and stereo disparity cues. The model also explicitly represents depth ordering and occlusion. Second, we introduce a stereoscopic dataset with frames extracted from feature-length movies “StreetDance 3D" and “Pina". The dataset contains 587 annotated human poses, 1158 bounding box annotations and 686 pixel-wise segmentations of people. The dataset is composed of indoor and outdoor scenes depicting multiple people with frequent occlusions. We demonstrate results on our new challenging dataset, as well as on the H2view dataset from (Sheasby et al. ACCV 2012).

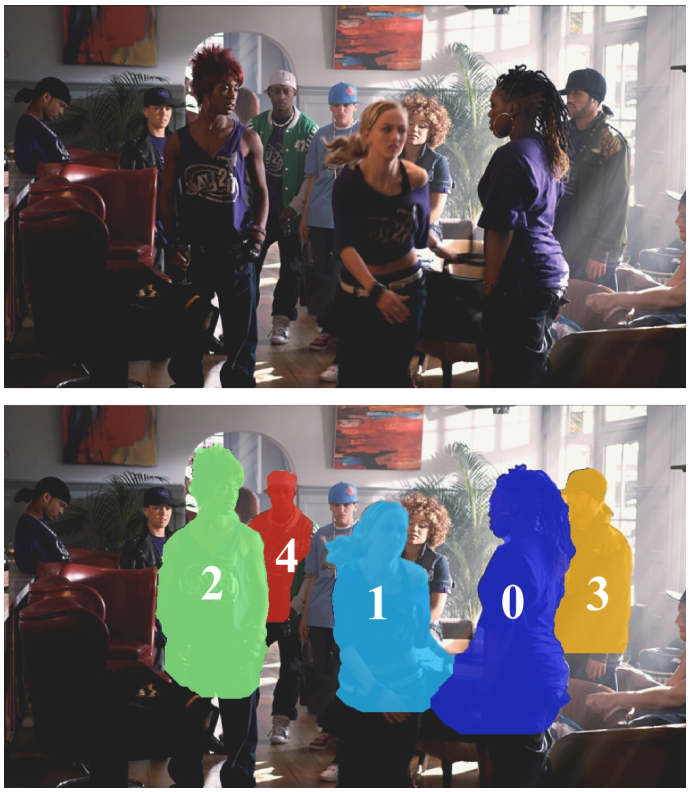

Figure 13. We segment multiple people in the scene, estimate their poses and relative front-to-back order, denoted by the numbers in the image below, in every frame of a video sequence.

7.3.9. Encoding Feature Maps of CNNs for Action Recognition

Participants: Xiaojiang Peng, Cordelia Schmid.

In [29] We describe our approach for action classification in the THUMOS Challenge 2015. Our approach is based on two types of features, improved dense trajectories and CNN features, as illustrated in Figure 14. For trajectory features, we extract HOG, HOF, MBHx, and MBHy descriptors and apply Fisher vector encoding. For CNN features, we utilize a recent deep CNN model, VGG19, to capture appearance features and use VLAD encoding to encode/pool convolutional feature maps which shows better performance than average pooling of feature maps and full-connected activation features.

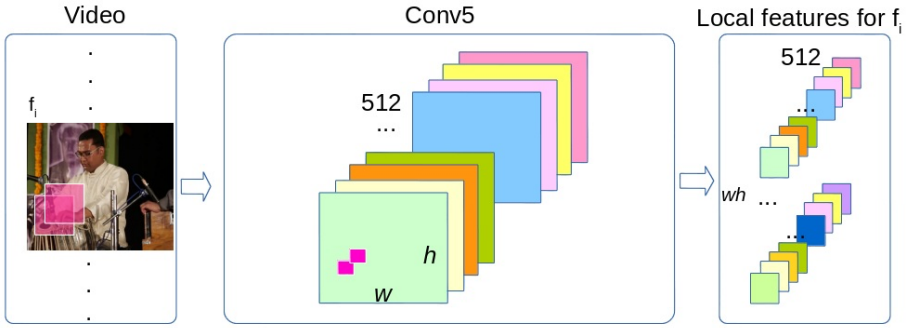

*Figure 14. Local features from convolutional feature maps. Each pixel (pink square in the middle image) in the Conv5 feature map is actually a feature for the corresponding patch in original frame. We obtain w*h 512-D features for frame f_i.*

7.3.10. Online Object Tracking with Proposal Selection

Participants: Yang Hua, Karteek Alahari, Cordelia Schmid.

Tracking-by-detection approaches are some of the most successful object trackers in recent years. Their success is largely determined by the detector model they learn initially and then update over time. However, under challenging conditions where an object can undergo transformations, e.g., severe rotation, these methods are found to be lacking. In [14], we address this problem by formulating it as a proposal selection task and making two contributions. The first one is introducing novel proposals estimated from the geometric transformations undergone by the object, and building a rich candidate set for predicting the object location. The second one is devising a novel selection strategy using multiple cues, i.e., detection score and edgeness score computed from state-of-the-art object edges and motion boundaries. We extensively evaluate our approach on the visual object tracking 2014 challenge and online tracking benchmark datasets, and show the best performance. Sample results are shown in Figure 15. Our tracker based on this method has recently won the visual object tracking challenge (VOT-TIR) organized as part of ICCV 2015 in Santiago, Chile.

8. Bilateral Contracts and Grants with Industry

8.1. MBDA

Participants: Jakob Verbeek, Julien Bardonnet.

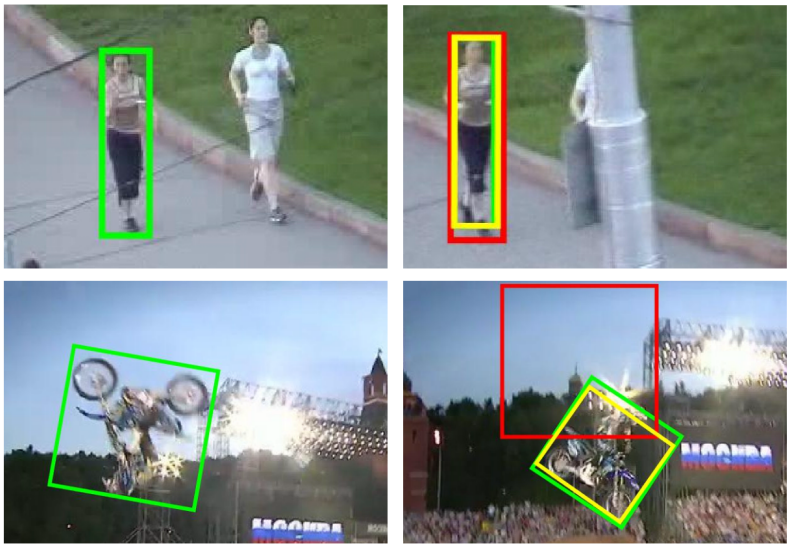

Figure 15. Sample frames (cropped) from the jogging (top row) and motocross (bottom row) sequences. The ground truth annotation (green) in the first frame (left) is used to train our tracker and the winner of VOT2014 challenge. We show these two tracking results (right) on another frame in the sequence. Our method (yellow) successfully tracks objects undergoing deformations unlike winner of VOT2014 challenge (red).

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

8.2. Google

Participants: Karteek Alahari, Cordelia Schmid.

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

8.3. Facebook

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

End of 2015 we received a gift from Facebook. The collaboration will start in 2016. The topics include image retrieval with CNN based descriptors, weakly supervised semantic segmentation, and learning structure models for action recognition in videos.

8.4. 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.

8.5. 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.

8.6. Xerox Research Center Europe

Participants: Zaid Harchaoui, Mattis Paulin, Karteek Alahari, Vladyslav Sydorov, 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 deep learning based image description and pose estimation in videos.

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. ANR Project Physionomie

Participants: Jakob Verbeek, Shreyas Saxena, Guosheng Hu.

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

9.1.2. 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.

9.1.3. MASTODONS Program CNRS - Project Titan

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: Jan 2015-Dec. 2015

9.1.4. 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.

9.2. European Initiatives

9.2.1. FP7 & H2020 Projects

9.2.1.1. 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 ended in May 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.

9.2.1.2. ERC Advanced grant Allegro

Participants: Cordelia Schmid, Karteek Alahari, Jerome Revaud, Pavel Tokmakov, Nicolas Chesneau, Vicky Kalogeiton, Konstantin Shmelkov, Daan Wynen, Xiaojiang Peng.

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

9.3. International Initiatives

9.3.1. 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. Yuansi Chen, from UC Berkeley visited LEAR in 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. J. Mairal also started a collaboration with Peter Richtarik, professor at university of Edinburgh and Dominik Csiba (PhD student), on the topic of local low-rank matrix estimation.
- **MPI Tübingen:** C. Schmid collaborates with M. Black, a research director at MPI since 2013. She spent one month at MPI in January 2015. End of 2015 she was award a Humbolt research award funding a long-term research project with colleagues at MPI.
- **Technion:** J. Mairal started a collaboration with Yonina Eldar (Technion) and Andreas Tillmann (Darmstadt university) to develop dictionary learning techniques for phase retrieval. Andreas Tillmann visited the LEAR team for a week in May 2015. Their collaboration resulted in a paper accepted to the ICASSP'16 conference.

9.3.2. 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 April 2016. The funds are meant to support scientific and scholarly exchanges and collaboration between the two teams.

9.4. International Research Visitors

9.4.1. Visits of International Scientists

Andreas Tillmann (Darmstadt university) and Dominik Csiba (Edinburgh university) visited Julien Mairal for a week, respectively in May and October 2015.

9.4.2. Visits to International Teams

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

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific events organisation

10.1.1.1. General chair, scientific chair

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

10.1.1.2. Member of the organizing committees

- C. Schmid and K. Alahari organized the ALLEGRO Workshop on weakly supervised learning and video recognition, Grenoble, 2015.
- J. Verbeek: Co-organizer Physionomie workshop at European Academy of Forensic Science conference, Prague, Czech Republic, September 9
- Z. Harchaoui: Co-organizer of the Optimization and Statistical Learning workshop, Les Houches, France, January 2015.
- Z. Harchaoui and J. Mairal: Co-organizers of the Large-scale Learning summer school, Grenoble, France, Spring 2015.
- Z. Harchaoui: Co-organizer of the Future of Artificial Intelligence Symposium, New York University, January 2016.
- G. Rodez: co-organizer of the IEEE CVPR Workshop on Observing and Understanding hands in action (HANDS 2015)

10.1.2. Scientific events selection

10.1.2.1. Member of the conference program committees

- C. Schmid: area chair for ICCV'15.
- J. Mairal: area chair for ICML'15, ICCV'15, CVPR'16, ICLR'16, NIPS'16.
- Z. Harchaoui: area chair for ICML'15, ICML'16, NIPS'16.
- J. Verbeek: area chair for CVPR'15. tutorial chair for ECCV'16.

10.1.2.2. Reviewer

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

10.1.3. Journal

10.1.3.1. Member of the editorial boards

- C. Schmid: Editor in Chief of the International Journal of Computer Vision, since 2013.
- C. Schmid: Associate editor for Foundations and Trends in Computer Graphics and Vision, since 2005.
- J. Verbeek: Associate editor for Image and Vision Computing Journal, since 2011.
- J. Verbeek: Associate editor for the International Journal on Computer Vision, since 2014.
- J. Mairal: Associate editor of the International Journal of Computer Vision (IJCV), since 2015.
- J. Mairal: Senior associate editor for IEEE Signal Processing Letters, since August 2014 (senior editor since Feb. 2015).
- J. Mairal: Associate editor of Journal on Mathematical Imaging and Vision (JMIV), since 2015.
- J. Mairal: Guest editor for the Special Issue on Sparse Coding of the International Journal of Computer Vision. 2015.
- K. Alahari. Guest editor for the Special Issue on "Higher Order Graphical Models in Computer Vision: Modelling, Inference & Learning" IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 37, Issue 7, July 2015

10.1.3.2. Reviewer - Reviewing activities

The permanent members of the team reviewed numerous papers for numerous international journals in computer vision (IJCV, PAMI,CVIU), machine learning (JMLR, Machine Learning). Some of them are also reviewing for journals in optimization (SIAM Journal on Optimization), image processing (SIAM Imaging Science), information theory (IEEE Transactions on Information Theory), statistics (Bernoulli, Annals of Statistics, Journal of American Statistical Association).

10.1.4. Invited talks

- C. Schmid: Speaker at the Scenes from Video Workshop, Santa Cruz, Chile, December 2015.
- C. Schmid: Invited speaker at the Google Deep Video Workshop, Santa Cruz, USA, November 2015.
- C. Schmid: Invited speaker at the Human Robot Interaction Workshop at UC Berkeley, November 2015.
- C. Schmid: Invited speaker at workshop on pose recovery, action recognition, and cultural event recognition, in conjunction with CVPR'15, June 2015.
- C. Schmid: Seminar at Berkeley University, November 2015.
- C. Schmid: Seminar at Stanford University, June 2015.
- C. Schmid: Seminar at Google, Mountain View, June 2015.
- C. Schmid: Seminar at Facebook AI Research lab, New York, May 2015.
- C. Schmid: Seminar at CMU, Pittsburgh, May 2015.
- C. Schmid: Seminar at Oxford University, March 2015.
- C. Schmid: Seminar at Gatsby Computational Neuroscience Unit, London, March 2015.
- C. Schmid: Seminar at MPI, Tübingen, February 2015.
- K. Alahari: Keynote talk at 5th Workshop on Algorithmic issues for Inference in Graphical Models (AIGM), Paris, France, September 2015
- K. Alahari: Invited talk at Universidad de Cordoba, Cordoba, Spain, May 2015
- K. Alahari: Invited talk at 36th Pattern Recognition and Computer Vision Colloquium, CTU, Prague, Czech Republic, April 2015
- Z. Harchaoui: Seminar at New-York university, March 2015.
- Z. Harchaoui: Invited talk at the Inria at Silicon valley workshop, May 2015.
- Z. Harchaoui: Invited talk at the workshop on "large-scale kernel learning", ICML, Lille, July 2015.
- J. Verbeek: DGA workshop on Big Data in Multimedia Information Processing, invited speaker, Paris, France, October 22.
- J. Verbeek: Physionomie workshop at European Academy of Forensic Science conference, speaker, Prague, Czech Republic, September 9.
- J. Verbeek: StatLearn workshop, invited speaker, April 13, 2015, Grenoble, France.
- J. Verbeek: Société Francaise de Statistique, Institut Henri Poincaré, Paris, France, October 23.
- J. Verbeek: Center for Machine Perception, Czech Technical University, Prague, Czech Republic, September 8.
- J. Verbeek: Dept. of Information Engineering and Computer Science, University of Trento, Italy, March 16.
- J. Verbeek: Computer Vision Center, Barcelona, Spain, February 13.
- J. Mairal: Invited talk at the BASP frontiers workshop, February 2015.
- J. Mairal: StatLab Seminar at Cambridge University, March 2015.
- J. Mairal: Invited talk at the CIMI workshop, Toulouse, October 2015.
- H. Lin: workshop TITAN, Grenoble, November 2015.

10.1.5. Scientific expertise

- C. Schmid is member of the PAMI-TC awards committee, and the PAMI-TC executive commitee.
- K. Alahari: Reviewer for the National Sciences and Engineering Research Council of Canada (NSERC).
- J. Mairal: reviewer for grant proposals from ANR and COFECUB.

10.1.6. Research administration

- C. Schmid is member of the "comité d'orientations scientifiques". Inria Grenoble, 2015.
- J. Mairal participated to the prospecting group from the PROSPER network in 2015.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Doctorat: C. Schmid, Course on action recogition at Computational Vision Summer School, Freudenstadt, Germany, July 2015.

Doctorat: Z. Harchaoui, "Machine Learning for Computer Vision," MLSS, Kyoto, Japan.

Doctorat: J. Mairal, "Dictionary Learning", 6H, CIMI summer school, Toulouse, France.

Doctorat: J. Mairal, "Introduction to sparse estimation", 4H, BigOptim summer school, Grenoble, France.

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

Master: Z. Harchaoui, "Computational Machine Learning", New York University.

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

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

Master: J. Mairal, "Introduction to sparse estimation", 4H, M2, PSL-ITI, France.

Master: M. Douze, K. Alahari, "Bases de donnees multimedia", Grenoble INP - ENSIMAG, January 2015.

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", 67.5H TD, L1, DLST Grenoble.

10.2.2. Supervision

PhD: D. Oneata, "Robust and efficient models for action recognition and localization", Grenoble Univ, July 2015, advisors: C. Schmid and J. Verbeek.

PhD: D. Potapov, Supervised Learning Approaches for Automatic Structuring of Videos, July 2015. Advisors: Z. Harchaoui and C. Schmid.

10.2.3. Juries

C. Schmid: Vincent Lepetit, decembre 2015, HDR examinateur, univ. Grenoble

C. Schmid: Nicolas Thome, juillet 2015, HDR examinateur, univ. Pierre et Marie Curie, Paris

C. Schmid: Vincent Delaitre, avril 2015, these, examinateur, ENS Ulm

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] D. ONEATA. *Robust and efficient models for action recognition and localization*, Université Grenoble Alpes, July 2015, https://tel.archives-ouvertes.fr/tel-01217362

[2] D. POTAPOV. *Supervised Learning Approaches for Automatic Structuring of Videos*, Université Grenoble Alpes, July 2015, https://tel.archives-ouvertes.fr/tel-01238100

Articles in International Peer-Reviewed Journals

[3] Z. AKATA, F. PERRONNIN, Z. HARCHAOUI, C. SCHMID. *Label-Embedding for Image Classification*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", September 2015, Accepted, 29.09.2015, https://hal.inria.fr/hal-01207145

[4] E. BERNARD, L. JACOB, J. MAIRAL, E. VIARA, J.-P. VERT. *A convex formulation for joint RNA isoform detection and quantification from multiple RNA-seq samples*, in "BMC Bioinformatics", 2015, vol. 16, n^{o} 1, 262 p. [*DOI :* 10.1186/s12859-015-0695-9], https://hal-mines-paristech.archives-ouvertes.fr/hal-01123141

[5] R. G. CINBIS, J. VERBEEK, C. SCHMID. *Approximate Fisher Kernels of non-iid Image Models for Image Categorization*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2015 [*DOI :* 10.1109/TPAMI.2015.2484342], https://hal.inria.fr/hal-01211201

[6] M. DOUZE, J. REVAUD, J. VERBEEK, H. JÉGOU, C. SCHMID. *Circulant temporal encoding for video retrieval and temporal alignment*, in "International Journal of Computer Vision", 2016, https://hal.inria.fr/hal-01162603

[7] J. MAIRAL. *Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning*, in "SIAM Journal on Optimization", April 2015, vol. 25, n^{o} 2, pp. 829–855 [*DOI :* 10.1137/140957639], https://hal.inria.fr/hal-00948338

[8] 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", August 2015, vol. 37, n^{o} 8, pp. 1643 - 1655 [*DOI :* 10.1109/TPAMI.2014.2369050], https://hal.inria.fr/hal-01089660

[9] H. WANG, D. ONEATA, J. VERBEEK, C. SCHMID. *A robust and efficient video representation for action recognition*, in "International Journal of Computer Vision", July 2015, pp. 1–20 [*DOI :* 10.1007/s11263-015-0846-5], https://hal.inria.fr/hal-01145834

International Conferences with Proceedings

[10] P. BOJANOWSKI, R. LAJUGIE, E. GRAVE, F. BACH, I. LAPTEV, J. PONCE, C. SCHMID. *Weakly-Supervised Alignment of Video With Text*, in "ICCV 2015 - IEEE International Conference on Computer Vision", Santiago, Chile, December 2015, https://hal.inria.fr/hal-01154523

[11] M. CHO, S. KWAK, C. SCHMID, J. PONCE. *Unsupervised Object Discovery and Localization in the Wild: Part-based Matching with Bottom-up Region Proposals*, in "CVPR 2015 - IEEE Conference on Computer Vision & Pattern Recognition", Boston, United States, June 2015, https://hal.inria.fr/hal-01110036

[12] G. CHÉRON, I. LAPTEV, C. SCHMID. *P-CNN: Pose-based CNN Features for Action Recognition*, in "ICCV 2015 - IEEE International Conference on Computer Vision", Santiago, Chile, December 2015, https://hal.inria.fr/hal-01187690

[13] Z. HARCHAOUI, A. JUDITSKY, A. NEMIROVSKI, D. OSTROVSKY. *Adaptive Recovery of Signals by Convex Optimization*, in "JMLR Workshop and Conference Proceedings", Paris, France, July 2015, vol. Proceedings of The 28th Conference on Learning Theory, n^{o} 40, https://hal.inria.fr/hal-01250215

[14] Y. HUA, K. ALAHARI, C. SCHMID. *Online Object Tracking with Proposal Selection*, in "ICCV 2015 - IEEE International Conference on Computer Vision", Santiago, Chile, IEEE, December 2015, https://hal.inria.fr/hal-01207196

[15] S. KWAK, M. CHO, I. LAPTEV, J. PONCE, C. SCHMID. *Unsupervised Object Discovery and Tracking in Video Collections*, in "ICCV 2015 - IEEE International Conference on Computer Vision", Santiago, Chile, December 2015, https://hal.archives-ouvertes.fr/hal-01153017

[16] H. LIN, J. MAIRAL, Z. HARCHAOUI. *A Universal Catalyst for First-Order Optimization*, in "Advances in Neural Information Processing Systems (NIPS)", Montreal, France, December 2015, main paper (9 pages) + appendix (21 pages), https://hal.inria.fr/hal-01160728

[17] M. PAULIN, M. DOUZE, Z. HARCHAOUI, J. MAIRAL, F. PERRONNIN, C. SCHMID. *Local Convolutional Features with Unsupervised Training for Image Retrieval*, in "IEEE International Conference on Computer Vision (ICCV)", Santiago, Chile, December 2015, https://hal.inria.fr/hal-01207966

[18] J. REVAUD, P. WEINZAEPFEL, Z. HARCHAOUI, C. SCHMID. *EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow*, in "CVPR 2015 - IEEE Conference on Computer Vision & Pattern Recognition", Boston, United States, June 2015, https://hal.inria.fr/hal-01142656

[19] G. ROGEZ, J. S. SUPANCIC, D. RAMANAN. *Understanding Everyday Hands in Action from RGB-D Images*, in "ICCV 2015 - IEEE International Conference on Computer Vision", Santiago, Chile, December 2015, https://hal.inria.fr/hal-01237011

[20] S. SAXENA, J. VERBEEK. *Coordinated Local Metric Learning*, in "ICCV ChaLearn Looking at People workshop", Santiago, Chile, December 2015, https://hal.inria.fr/hal-01215272

[21] J. S. SUPANCIC, G. ROGEZ, Y. YANG, J. SHOTTON, D. RAMANAN. *Depth-based hand pose estimation: data, methods, and challenges*, in "ICCV 2015 - IEEE International Conference on Computer Vision", Santiago, Chile, December 2015, https://hal.inria.fr/hal-01237023

[22] P. WEINZAEPFEL, Z. HARCHAOUI, C. SCHMID. *Learning to track for spatio-temporal action localization*, in "ICCV 2015 - IEEE International Conference on Computer Vision", Santiago, Chile, December 2015, https://hal.inria.fr/hal-01159941

[23] P. WEINZAEPFEL, J. REVAUD, Z. HARCHAOUI, C. SCHMID. *Learning to Detect Motion Boundaries*, in "CVPR 2015 - IEEE Conference on Computer Vision & Pattern Recognition", Boston, United States, June 2015, https://hal.inria.fr/hal-01142653

[24] V. ZADRIJA, J. KRAPAC, J. VERBEEK, S. ŠEGVIĆ. *Patch-level spatial layout for classification and weakly supervised localization*, in "German Conference on Pattern Recognition", Aachen, Germany, October 2015, https://hal.inria.fr/hal-01186677

Research Reports

[25] G. SHARMA, F. JURIE, C. SCHMID. *Expanded Parts Model for Semantic Description of Humans in Still Images*, Max-Planck Institute for Informatics ; GREYC CNRS UMR 6072, Universite de Caen ; Inria Grenoble - Rhône-Alpes, September 2015, https://hal.inria.fr/hal-01199160

Other Publications

[26] R. G. CINBIS, J. VERBEEK, C. SCHMID. *Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning*, September 2015, working paper or preprint, https://hal.inria.fr/hal-01123482

[27] B. HAM, M. CHO, C. SCHMID, J. PONCE. *Proposal Flow*, December 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01240281

[28] N. HE, Z. HARCHAOUI. *Semi-proximal Mirror-Prox for Nonsmooth Composite Minimization*, June 2015, working paper or preprint, https://hal.inria.fr/hal-01171567

[29] X. PENG, C. SCHMID. *Encoding Feature Maps of CNNs for Action Recognition*, 2015, CVPR'15 International Workshop and Competition on Action Recognition with a Large Number of Classes, https://hal.inria.fr/hal-01236843

[30] D. POTAPOV, M. DOUZE, J. REVAUD, Z. HARCHAOUI, C. SCHMID. *Beat-Event Detection in Action Movie Franchises*, August 2015, working paper or preprint, https://hal.inria.fr/hal-01183588

[31] J. REVAUD, P. WEINZAEPFEL, Z. HARCHAOUI, C. SCHMID. *DeepMatching: Hierarchical Deformable Dense Matching*, October 2015, working paper or preprint, https://hal.inria.fr/hal-01148432

[32] J. REVAUD, P. WEINZAEPFEL, Z. HARCHAOUI, C. SCHMID. *EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow*, April 2015, working paper or preprint, https://hal.inria.fr/hal-01097477

[33] G. VAROL, I. LAPTEV, C. SCHMID. *Long-term Temporal Convolutions for Action Recognition*, December 2015, working paper or preprint, https://hal.inria.fr/hal-01241518