Activity Report 2012

Project-Team LEAR

Learning and recognition in vision

Table of contents

Project-Team LEAR

Keywords: Computer Vision, Machine Learning, Video, Recognition

Beginning of the Team: 2003-07-01, End of the Team: 2015-12-31.

1. Members

Research Scientists

Cordelia Schmid [Team Leader, INRIA Research Director, DR1, HdR]
Zaid Harchaoui [INRIA Researcher, CR2 until August '12, CR1 starting November '12]
Julien Mairal [INRIA Researcher, CR2, October '12 – September '15]
Jakob Verbeek [INRIA Researcher, CR1]

Faculty Member

Roger Mohr [Professor émérite at ENSIMAG, HdR]

External Collaborators

Frédéric Jurie [Professor at University of Caen, HdR]
Laurent Zwald [Associate professor at UJF, LJK-SMS]

Engineers

Mohamed Ayari [QUAERO project, November '10 – April '12]
Matthijs Douze [INRIA engineer SED, 40%]
Guillaume Fortier [ITI Visages project, Qcompere project, October '10 – June '13]
Franck Thollard [QUAERO project, October '12 – October '13]

PhD Students

Zeynep Akata [Univ. Grenoble, Cifre grant XRCE, January '11 – January '14]
Ramazan Cinbis [Univ. Grenoble, INRIA PhD Scholarship, October '10 – October '13]
Adrien Gaidon [Univ. Grenoble, Microsoft/INRIA project, Oct. '08 – Oct. '12, PostDoc Nov. '12 – Jan. '13]
Yang Hua [Univ. Grenoble, Microsoft/INRIA project, October '12 – October '15]
Thomas Mensink [Univ. Grenoble, EU project CLASS Feb.'09–Sep.'09, Cifre grant XRCE Oct.'09–Oct.'12]
Dan Oneata [Univ. Grenoble, EU project AXES, QUAERO project, October '11 – October '14]
Federico Pierucci [Univ. Grenoble, Intern January '12 – September '12, PhD October '12 – September '15]
Danila Potapov [Univ. Grenoble, EU project AXES, QUAERO project, September '11 – August '14]
Alessandro Prest [ETH Zürich, QUAERO project, co-supervision with V. Ferrari, June '09 – July '12]
Gaurav Sharma [University of Caen, ANR project SCARFACE, co-superv. with F. Jurie, Oct. '09 – Oct. '12]
Philippe Weinzaepfel [Univ. Grenoble, Intern Jan. '12 – Sep. '12, PhD Oct. '12 – Oct. '15]

Post-Doctoral Fellows

Ahmed Gamal-Eldin [QUAERO project, January '12 – December '12]
Anoop Cherian [Microsoft/INRIA project, November '12 – August '13]
Albert Gordo [MBDA project, August '12 – January '14]
Jerome Revaud [GAIA project, QUAERO project, June '11 – May '13]
Heng Wang [QUAERO project, July '12 – June '13]

Administrative Assistants

Florence Polge [Secretary INRIA, September 2011 – March 2012]
Assia Hraoubia [Secretary INRIA, April 2012 – September 2012]
Nathalie Gillot [Secretary INRIA, since October 2012]

Other

Michael Guerzhoy [Intern, MBDA project, March '12 – August '13]

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

- **Excellent results at TrecVid MED.** This year we participated for the second time in the Multimedia Event Detection (MED) track of TrecVid, one of the major benchmarks in automatic video analysis. In this task 25 event categories (from "making a sandwich" to "attempting a bicycle trick") have to be detected in a video corpus of 4,000 hours. We ranked first out of 13 participants on the ad-hoc event category task, and 2-nd out of 17 participants for the pre-specified event category task.

- **ERC advanced grant.** In 2012 Cordelia Schmid was awarded an ERC advanced grant for the ALLEGRO project on Active Large-scale LEarninG for visual RecOgnition. The aim of ALLEGRO is to automatically learn from large quantities of data with weak labels. In 2012 C. Schmid was also nominated IEEE fellow.
- **INRIA Visual Recognition and Machine Learning Summer School**. This year we co-organized the third edition of the INRIA Visual Recognition and Machine Learning Summer School in Grenoble. It attracted a total of 182 participants (48 from France, 94 from Europe and 40 from America and Asia).

3. Scientific Foundations

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 cross-modal retrieval of images given text queries, and vice-versa. 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

---

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

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. Transfer of such techniques is the goal of the start-up MilPix, to which our current technologies for image search are licensed. In a collaboration with Technosens we transfer face recognition technology, which they exploit to identify users of a system and adapt the interface to the user.

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

5.1. Face recognition

Participants: Guillaume Fortier [correspondant], Jakob Verbeek.

In a collaboration with Technosens (a start-up based in Grenoble) we are developing an efficient face recognition library. During 18 months Guillaume Fortier, financed by INRIA's technology transfer program, had streamlined code developed by different former team members on various platforms. This encompasses detection of characteristic points on the face (eyes, nose, mouth), computing appearance features on these points, and learning metrics on the face descriptors that are useful for face verification (faces of the same person are close, faces of different people are far away). See http://lear.inrialpes.fr/~fortier/software.php.

5.2. 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 to reproduce the results from [19] 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.

5.3. 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 [6], [16] and in our submisssion to the Trecvid 2012 MED task [31]. 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.

5.4. Video descriptors

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

We have developed and made on-line available software for video description based on dense trajectories and motion boundary histograms [28]. 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 scalability to large datasets. On the one hand we explored the effect of sub-sampling the video input both spatially and temporally, and evaluated the impact on the quality of the descriptors. On the other hand we avoid writing the raw MBH descriptors to disk, but rather aggregate them directly into a signature for the complete video using Fisher vectors, or bag-of-word descriptors. This allowed us to use these descriptors on the 4,000 hour video dataset of the TrecVid 2012 MED task.

6. New Results

6.1. Visual recognition in images

6.1.1. *Correlation-Based Burstiness for Logo Retrieval*

Participants: Matthijs Douze, Jerome Revaud, Cordelia Schmid.

Detecting logos in photos is challenging. A reason is that logos locally resemble patterns frequently seen in random images. In [21] we propose to learn a statistical model for the distribution of incorrect detections output by an image matching algorithm. It results in a novel scoring criterion in which the weight of correlated keypoint matches is reduced, penalizing irrelevant logo detections. In experiments on two very different logo retrieval benchmarks, our approach largely improves over the standard matching criterion as well as other state-of-the-art approaches.

Figure 1. Illustration of a logo detected by our method.

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

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

In [19] we propose a benchmark of several objective functions for large-scale image classification: we compare the one-vs-rest, multiclass, ranking and weighted average ranking SVMs. Using stochastic gradient descent optimization, we can scale the learning to millions of images and thousands of classes. Our experimental evaluation shows that ranking based algorithms do not outperform a one-vs-rest strategy and that the gap between the different algorithms reduces in case of high-dimensional data. We also show that for one-vs-rest, learning through cross-validation the optimal degree of imbalance between the positive and the negative samples can have a significant impact. Furthermore, early stopping can be used as an effective regularization strategy when training with stochastic gradient algorithms. Following these "good practices", we were able to improve the state-of-the-art on a large subset of 10K classes and 9M of images of ImageNet from 16.7% accuracy to 19.1%. Some qualitative results can be seen in Figure 2.
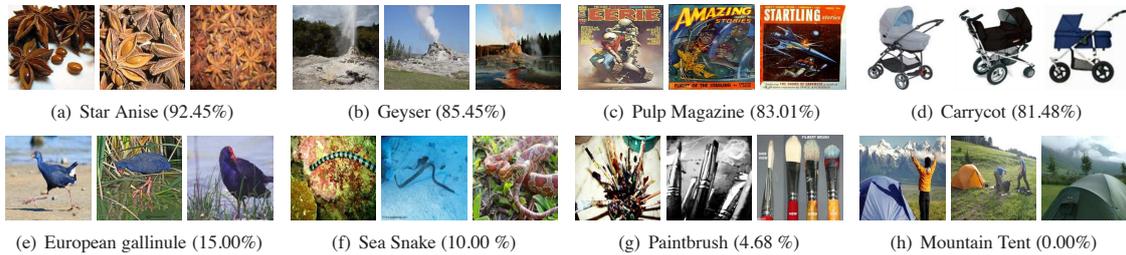
(a) Star Anise (92.45%)   (b) Geyser (85.45%)   (c) Pulp Magazine (83.01%)   (d) Carrycot (81.48%)

(e) European gallinule (15.00%)   (f) Sea Snake (10.00 %)   (g) Paintbrush (4.68 %)   (h) Mountain Tent (0.00%)

Figure 2. ImageNet10K results (top-1 accuracy in %) obtained with w-OVR and 130K-dim Fisher vectors. (a-d) Sample classes among the best performing ones. (e-h) Sample classes among the worst performing ones.

6.1.3. Discriminative Spatial Saliency for Image Classification

Participants: Frédéric Jurie [Université de Caen], Cordelia Schmid, Gaurav Sharma.

In many visual classification tasks the spatial distribution of discriminative information is (i) non uniform e.g. "person reading" can be distinguished from "taking a photo" based on the area around the arms i.e. ignoring the legs, and (ii) has intra class variations e.g. different readers may hold the books differently. Motivated by these observations, we propose in [22] to learn the discriminative spatial saliency of images while simultaneously learning a max-margin classifier for a given visual classification task. Using the saliency maps to weight the corresponding visual features improves the discriminative power of the image representation. We treat the saliency maps as latent variables and allow them to adapt to the image content to maximize the classification score, while regularizing the change in the saliency maps. See Figure 3 for an illustration. Our experimental results on three challenging datasets, for (i) human action classification, (ii) fine grained classification, and (iii) scene classification, demonstrate the effectiveness and wide applicability of the method.
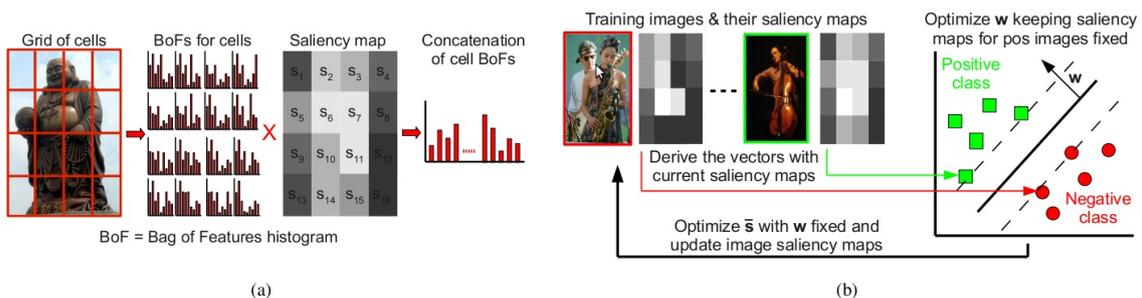
Figure 3. (a) The images are represented by concatenation of cell bag-of-features weighted by the image saliency maps. (b) We propose to use a block coordinate descent algorithm for learning our model. As in a latent SVM, we optimize in one step the weight vector w keeping the saliency maps of the positive images fixed, and in the other step we optimize the saliency keeping w fixed.

6.1.4. Tree-structured CRF Models for Interactive Image Labeling

Participants: Gabriela Csurka [XRCE], Thomas Mensink, Jakob Verbeek.

In [8] we propose structured prediction models for image labeling that explicitly take into account dependencies among image labels. In our tree structured models, image labels are nodes, and edges encode dependency relations. To allow for more complex dependencies, we combine labels in a single node, and use mixtures of trees. Our models are more expressive than independent predictors, and lead to more accurate label predictions. The gain becomes more significant in an interactive scenario where a user provides the value of some of the image labels at test time. Such an interactive scenario offers an interesting trade-off between label accuracy and manual labeling effort. The structured models are used to decide which labels should be set by the user, and transfer the user input to more accurate predictions on other image labels. We also apply our models to attribute-based image classification, where attribute predictions of a test image are mapped to class probabilities by means of a given attribute-class mapping. Experimental results on three publicly available benchmark data sets show that in all scenarios our structured models lead to more accurate predictions, and leverage user input much more effectively than state-of-the-art independent models.

6.1.5. *Metric Learning for Large Scale Image Classification: Generalizing to new classes at near-zero cost*

Participants: Gabriela Csurka [XRCE], Thomas Mensink, Florent Perronnin [XRCE], Jakob Verbeek.

In [18], [27] we consider the task of large scale image classification in open ended datasets. Many real-life datasets are open-ended and dynamic: new images are continuously added to existing classes, new classes appear over time and the semantics of existing classes might evolve too. In order to be able to handle new images and new classes at near-zero cost we consider two distance based classifiers, the k-nearest neighbor (k-NN) and nearest class mean (NCM) classifiers. For the NCM classifier we introduce a new metric learning approach, which has advantageous properties over the classical Fisher Discriminant Analysis. We also introduce an extension of the NCM classifier to allow for richer class representations, using multiple centroids per class. Experiments on the ImageNet 2010 challenge dataset, which contains over one million training images of thousand classes, show that, surprisingly, the NCM classifier compares favorably to the more flexible k-NN classifier. Moreover, the NCM performance is comparable to that of linear SVMs which obtain current state-of-the-art performance. Experimentally we study the generalization performance to classes that were not used to learn the metrics. Using a metric learned on 1,000 classes, we show results for the ImageNet-10K dataset which contains 10,000 classes, and obtain performance that is competitive with the current state-of-the-art, while being orders of magnitude faster. Furthermore, we show how a zero-shot class prior based on the ImageNet hierarchy can improve performance when few training images are available. See Figure 4 for an illustration.

6.2. Learning and statistical models

6.2.1. *Image categorization using Fisher kernels of non-iid image models*

Participants: Ramazan Cinbis, Cordelia Schmid, Jakob Verbeek.

Bag of visual words treat images as an orderless sets of local regions and represent them by visual word frequency histograms. Implicitly, regions are assumed to be identically and independently distributed (iid), which is a very poor assumption from a modeling perspective; see Figure 5 for an illustration. In [13], we introduce non-iid models by treating the parameters of bag-of-word models as latent variables which are integrated out, rendering all local regions dependent. Using the Fisher kernel we encode an image by the gradient of the data log-likelihood with respect to hyper-parameters that control priors on the model parameters. In fact, our models naturally generate transformations similar to taking square-roots, providing an explanation of why such non-linear transformations have proven successful. Using variational inference we extend the basic model to include Gaussian mixtures over local descriptors, and latent topic models to capture the co-occurrence structure of visual words, both improving performance. Our models yields state-of-the-art image categorization performance using linear classifiers, without using non-linear kernels, or (approximate) explicit embeddings thereof, e.g. by taking the square-root of the features.

	horseshoe crab 0.99%	African elephant 0.99%	mongoose 0.94%	Indian elephant 0.88%	dingo 0.87%	L2
Cliff dwelling L2 11.0% - Mah. 99.9%	cliff 0.07%	dam 0.00%	stone wall 0.00%	brick 0.00%	castle 0.00%	Mah.
	shopping cart 1.07%	unicycle 0.84%	covered wagon 0.83%	garbage truck 0.79%	forklift 0.78%	L2
Gondola L2 4.4% - Mah. 99.7%	dock 0.11%	canoe 0.03%	fishing rod 0.01%	bridge 0.01%	boathouse 0.01%	Mah.
	crane 0.87%	stupa 0.83%	roller coaster 0.79%	bell cote 0.78%	flagpole 0.75%	L2
Palm L2 6.4% - Mah. 98.1%	cabbage tree 0.81%	pine 0.30%	pandanus 0.14%	iron tree 0.07%	logwood 0.06%	Mah.

Figure 4. Examples of three classes, and the five most similar classes for each according to the standard ℓ_2 metric and our learned Mahalanobis metric.

Figure 5. Illustration of why local image patches are not independent: we can easily guess the image content in the masked areas.

6.2.2. *Conditional gradient algorithms for machine learning*
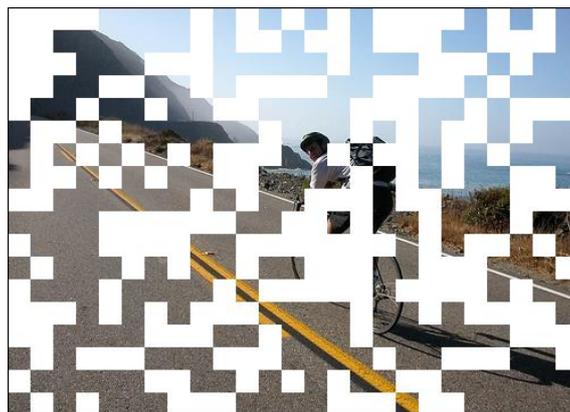
Participants: Zaid Harchaoui, Anatoli Juditsky [UJF], Arkadi Nemirovski [Georgia Tech].

In [17] we consider convex optimization problems arising in machine learning in high-dimensional settings. For several important learning problems, such as e.g. noisy matrix completion, state-of-the-art optimization approaches such as composite minimization algorithms are difficult to apply and do not scale up to large datasets. We study three conditional gradient-type algorithms, suitable for large-scale problems, and derive their finite-time convergence guarantees. Promising experimental results are presented on two large-scale real-world datasets.

6.2.3. *Large-scale classification with trace-norm regularization*

Participants: Matthijs Douze, Miro Dudik [Microsoft Research], Zaid Harchaoui, Jérôme Malick [BiPoP Team INRIA Grenoble], Mattis Paulin [ETHZ].

In [16] we introduce a new scalable learning algorithm for large-scale multi-class image classification, based on the multinomial logistic loss and the trace-norm regularization penalty. Reframing the challenging non-smooth optimization problem into a surrogate infinite-dimensional optimization problem with a regular ℓ_1-regularization penalty, we propose a simple and provably efficient accelerated coordinate descent algorithm. Furthermore, we show how to perform efficient matrix computations in the compressed domain for quantized dense visual features, scaling up to 100,000s examples, 1,000s-dimensional features, and 100s of categories. Promising experimental results on the "Fungus", "Ungulate", and "Vehicles" subsets of ImageNet are presented, where we show that our approach performs significantly better than state-of-the-art approaches for Fisher vectors with 16 Gaussians.

6.2.4. *Tree-walk kernels for computer vision*

Participants: Francis Bach [INRIA SIERRA team], Zaid Harchaoui.

In [25] we propose a family of positive-definite kernels between images, allowing to compute image similarity measures respectively in terms of color and of shape. The kernels consists in matching subtree-patterns called "tree-walks" of graphs extracted from the images, e.g. the segmentation graphs for color similarity and graphs of the discretized shapes or the point clouds in general for shape similarity. In both cases, we are able to design computationally efficient kernels which can be computed in polynomial-time in the size of the graphs, by leveraging specific properties of the graphs at hand such as planarity for segmentation graphs or factorizability of the associated graphical model for point clouds. Our kernels can be used by any kernel-based learning method, and hence we present experimental results for supervised and semi-supervised classification as well as clustering of natural images and supervised classification of handwritten digits and Chinese characters from few training examples.

6.2.5. *Lifted coordinate descent for learning with trace-norm regularization*

Participants: Miro Dudik [Microsoft Research], Zaid Harchaoui, Jérôme Malick [BiPoP Team INRIA Grenoble].

In [14] we consider the minimization of a smooth loss with trace-norm regularization, which is a natural objective in multi-class and multi-task learning. Even though the problem is convex, existing approaches rely on optimizing a non-convex variational bound, which is not guaranteed to converge, or repeatedly perform singular-value decomposition, which prevents scaling beyond moderate matrix sizes. We lift the non-smooth convex problem into an infinitely dimensional smooth problem and apply coordinate descent to solve it. We prove that our approach converges to the optimum, and is competitive or outperforms the state of the art.

6.3. Recognition in video

6.3.1. *Large-scale multi-media event detection in video*

Participants: Matthijs Douze, Zaid Harchaoui, Dan Oneata, Danila Potapov, Jerome Revaud, Cordelia Schmid, Jochen Schwenninger [Fraunhofer Institute, Bonn], Jakob Verbeek, Heng Wang.

This year we participated in the TrecVid Multimedia Event Detection (MED) task. The goal is to detect events categories (such as "birthday party", or "changing a vehicle tire") in a large collection of around 100,000 videos with a total duration of around 4,000 hours. To this end we implemented an efficient system based on our recently developed MBH video descriptor (see Section 5.4) , SIFT descriptors and, MFCC audio descriptors (contributed by Fraunhofer Institute). All these low-level descriptors are encoded using the Fisher vector representation (see Section 5.3). In addition we implemented an optical character recognition (OCR) system to extract textual features from the video. The system is described in a forthcoming paper [31], and ranked first and second in two evaluations among the 17 systems submitted by different international teams participating to the task. See Figure 6 for an illustration.

Figure 6. Illustration of videos retrieved for two event categories. From left to right, we show for each a frame from (i) the top ranked video, (ii,iii) the first negative video, and the postive just before, and (iv) the last positive video.

6.3.2. *Learning Object Class Detectors from Weakly Annotated Video*
Participants: Javier Civera, Vittorio Ferrari, Christian Leistner, Alessandro Prest, Cordelia Schmid.

Object detectors are typically trained on a large set of still images annotated by bounding-boxes. In [20] we introduce an approach for learning object detectors from real-world web videos known only to contain objects of a target class. We propose a fully automatic pipeline that localizes objects in a set of videos of the class and learns a detector for it. The approach extracts candidate spatio-temporal tubes based on motion segmentation and then selects one tube per video jointly over all videos. See Figure 7 for an illustration. To compare to the state of the art, we test our detector on still images, i.e., Pascal VOC 2007. We observe that frames extracted from web videos can differ significantly in terms of quality to still images taken by a good camera. Thus, we formulate the learning from videos as a domain adaptation task. We show that training from a combination of weakly annotated videos and fully annotated still images using domain adaptation improves the performance of a detector trained from still images alone.

6.3.3. *Recognizing activities with cluster-trees of tracklets*
Participants: Adrien Gaidon, Zaid Harchaoui, Cordelia Schmid.

In [15] we address the problem of recognizing complex activities, such as pole vaulting, which are characterized by the composition of a large and variable number of different spatio-temporal parts. We represent a video as a hierarchy of mid-level motion components. This hierarchy is a data-driven decomposition specific to each video. We introduce a divisive clustering algorithm that can efficiently extract a hierarchy over a large
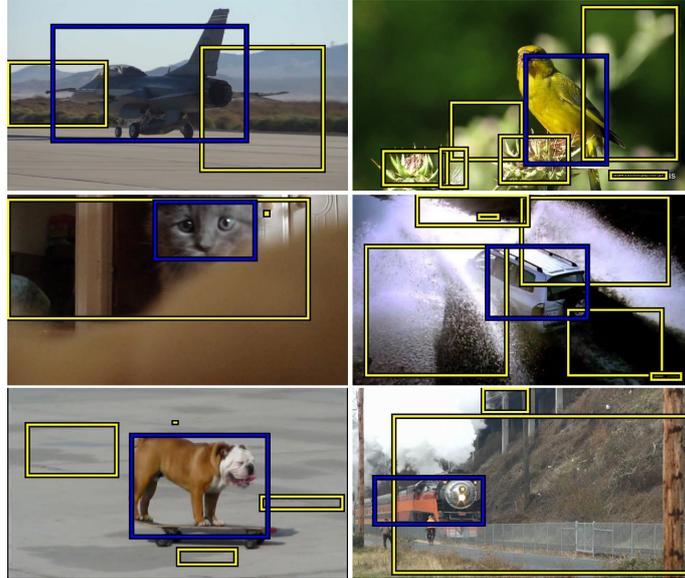
Figure 7. Yellow boxes represent tubes extracted by our method on the YouTube-Objects dataset. Blue boxes indicate the automatically selected tubes.

number of local trajectories. We use this structure to represent a video as an unordered binary tree. This tree is modeled by nested histograms of local motion features, see Figure 8. We provide an efficient positive definite kernel that computes the structural and visual similarity of two tree decompositions by relying on models of their edges. Contrary to most approaches based on action decompositions, we propose to use the full hierarchical action structure instead of selecting a small fixed number of parts. We present experimental results on two recent challenging benchmarks that focus on complex activities and show that our kernel on per-video hierarchies allows to efficiently discriminate between complex activities sharing common action parts. Our approach improves over the state of the art, including unstructured activity models, baselines using other motion decomposition algorithms, graph matching, and latent models explicitly selecting a fixed number of parts.

6.3.4. Action Detection with Actom Sequence Models

Participants: Adrien Gaidon, Zaid Harchaoui, Cordelia Schmid.

We address the problem of detecting actions, such as drinking or opening a door, in hours of challenging video data. In [26] 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 the temporal structure of actions as a sequence of histograms of actom-anchored visual features, see Figure 9 for an illutration. Our representation, which can be seen as a temporally structured extension of the bag-of-features, is flexible, sparse, and discriminative. Training requires the annotation of actoms for action examples. At test time, actoms are detected 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 temporal action detection: "Coffee and Cigarettes" and the "DLSB" dataset. We also adapt our approach to a classification by detection 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 detection, as well as baselines that detect actions with a sliding window method combined with bag-of-features.
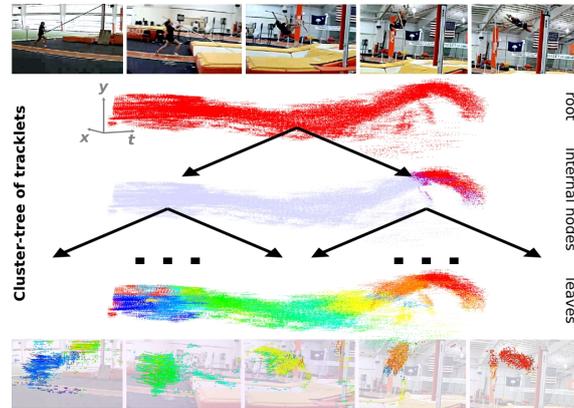
Figure 8. Illustration of tracklets found in a video and their hierarchical decomposition.
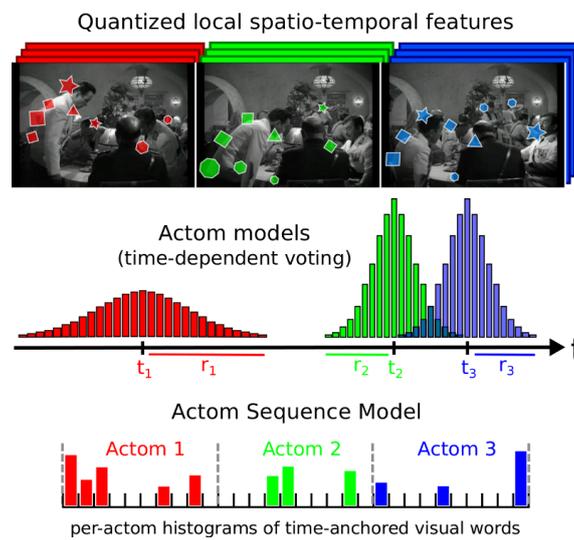

Figure 9. Illustration of the "Actom" video representation, see text for details.

6.3.5. *Action recognition by dense trajectories*

Participants: Alexander Kläser, Cheng-Lin Liu [Chinese Academy of Sciences], Cordelia Schmid, Heng Wang [Chinese Academy of Sciences].

In [28] we introduce 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 histograms (MBH) (see the visualization in Figure 10), 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.
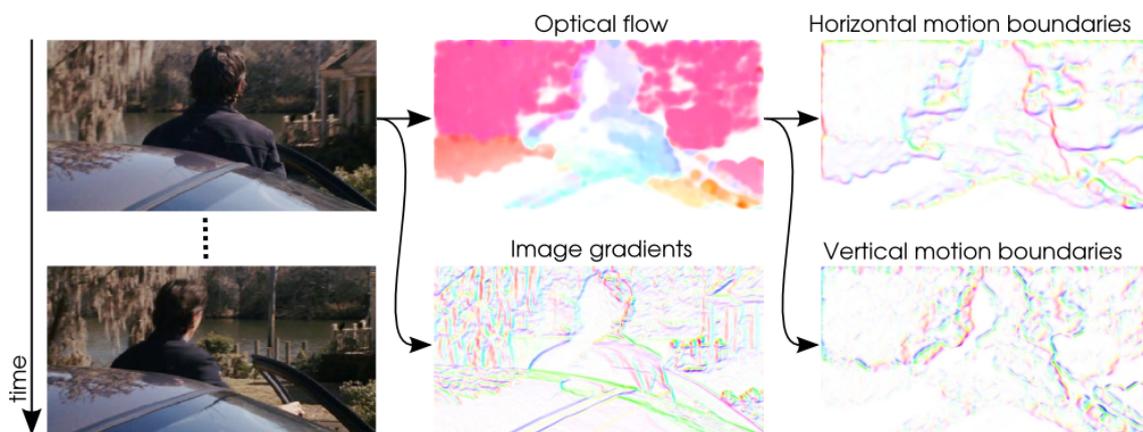
Figure 10. Illustration of the information captured by HOG, HOF, and MBH descriptors. Gradient/flow orientation is indicated by color (hue) and magnitude by saturation. The optical flow (top, middle) shows constant motion in the background, which is due to the camera movements. The motion boundaries (right) encode the relative motion between the person and the background.

7. Bilateral Contracts and Grants with Industry

7.1. Start-up Milpix

Participants: Hervé Jégou [INRIA Rennes], Cordelia Schmid.

In 2007, the start-up company MILPIX has been created by a former PhD student of the LEAR team, Christopher Bourez. The start-up exploits the technology developed by the LEAR team. Its focus is on large-scale indexing of images for industrial applications. Two software libraries were licensed to the start-up: BIGIMBAZ and OBSIDIAN.

7.2. MBDA Aerospatiale

Participants: Albert Gordo, Michael Guerzhoy, Frédéric Jurie [University of Caen], Cordelia Schmid.

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). In September 2010 started a new three-year contract on object localization and pose estimation.

7.3. MSR-INRIA joint lab: scientific image and video mining

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

This collaborative project, starting 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. Adrien Gaidon was funded by this project.

7.4. Xerox Research Center Europe

Participants: Zeynep Akata, Zaid Harchaoui, Thomas Mensink, Cordelia Schmid, Jakob Verbeek.

In a collaborative project with Xerox, staring October 2009, we work on cross-modal information retrieval. The challenge is to perform information retrieval and document classification in databases that contain documents in different modalities, such as texts, images, or videos, and documents that contain a combination of these. The PhD student Thomas Mensink was supported by a CIFRE grant obtained from the ANRT for the period 10/09 – 09/12. A second three-year collaborative project on large scale visual recognition started in 2011. The PhD student Zeynep Akata is supported by a CIFRE grant obtained from the ANRT for the period 01/11 – 01/14.

7.5. Technosens

Participants: Guillaume Fortier, Cordelia Schmid, Jakob Verbeek.

In October 2010 we started an 18 month collaboration with Technosens (a start-up based in Grenoble) in applying robust face recognition for application in personalized user interfaces. During 18 months an engineer financed by INRIA's technology transfer program, implemented and evaluated our face recognition system on Technosens hardware. Additional development aimed at dealing with hard real-world conditions.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. QUAERO Project
Participants: Mohamed Ayari, Matthijs Douze, Dan Oneata, Danila Potapov, Alessandro Prest, 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.

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.

8.2. European Initiatives

8.2.1. FP7 European Project AXES

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

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, Thomas Mensink, 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.

8.2.3. ERC Advanced grant Allegro

Participant: Cordelia Schmid.

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

8.3. International Initiatives

8.3.1. INRIA Associate Teams

- **Hyperion**: Large-scale statistical learning for visual recognition, 2012–2014
  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. In this associate team we collaborate with two teams of University of California Berkeley, headed respectively by Prof. Jitendra Malik and Prof. Nourredine El Karoui. It will allow 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

- **Microsoft Research NY:** Zaid Harchaoui has been collaborating since the fall 2010 with Miro Dudik, formerly from Yahoo! Research (until Spring 2012), and now in the recently setup Microsoft Research New York lab, on lifted coordinate descent algorithms for large-scale learning. This collaboration lead to several published papers, including an oral presentation at CVPR 2012. Zaid Harchaoui has visited Microsoft Research NY for one week in the fall 2012. We intend to pursue this fruitful collaboration in the coming years.

- **UC Berkeley:** This collaboration between Bin Yu, Jack Gallant, Yuval Benjamini (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, involving a student from UC Berkeley working on the extension of the current image model to videos.

- **UC Berkeley, Institut Curie:** In a collaboration between Jean-Philippe Vert, Elsa Bernard (Institut Curie), Laurent Jacob (UC Berkeley) and Julien Mairal (INRIA LEAR) we aim to develop novel efficient optimization techniques for identification and quantification of isoforms from RNA-Seq data. Elsa Bernard was a master student between April and August 2012. She was co-advised by Jean-Philippe Vert, Laurent Jacob and Julien Mairal. Elsa Bernard has now started her PhD at Institut Curie and the collaboration is still ongoing.

- **ETH Zürich:** We collaborate with V. Ferrari, junior professor at ETH Zürich, and recently appointed as assistant professor at University of Edinburgh. V. Ferrari and C. Schmid co-supervised a PhD student (A. Prest) on the subject of automatic learning of objects in images and videos [3], [9], [20]. A. Prest was bi-localized between ETH Zürich and INRIA Grenoble.

8.3.3. Participation In International Programs

- **France-Berkeley fund:** The LEAR team was awarded 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. The funds are meant to support scientific and scholarly exchanges and collaboration between the two teams.

9. Dissemination

9.1. Scientific Animation

- Conference, workshop, and summer school organization:
  - Z. Harchaoui: Co-organizer of the ICML 2012 Workshop on New Trends in RKHS and kernel-based methods, July 2012.

- – Z. Harchaoui: Co-organizer of the Optimization and Statistical Learning workshop, Les Houches, January 2013.
- – C. Schmid: Co-organizer of the INRIA Visual Recognition and Machine Learning Summer School, Grenoble, July 2012. Attracted a total of 182 participants (48 from France, 94 from Europe and 40 from America and Asia).
- – C. Schmid: Co-organizer IPAM workshop Large Scale Multimedia Search, January 9–13, 2012.
- Editorial boards:
  - – C. Schmid: International Journal of Computer Vision, since 2004.
  - – 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.
- Program chair:
  - – C. Schmid: ECCV '12.
- Area chair:
  - – J. Verbeek: BMVC '12, ECCV '12
  - – C. Schmid: NIPS '12, CVPR '13, ICCV '13
- Program committees:
  - – AISTATS 2012: Z. Harchaoui.
  - – BMVC 2012: T. Mensink.
  - – CVPR 2012: Z. Harchaoui, T. Mensink, C. Schmid, J. Verbeek.
  - – ECCV 2012: M. Douze, T. Mensink, J. Verbeek.
  - – ICML 2013: Z. Harchaoui, J. Mairal.
  - – NIPS 2012: Z. Harchaoui, J. Verbeek.
  - – WACV 2013: R. Cinbis.
  - – NIPS computational biology workshop 2012: J. Mairal.
- Prizes:
  - – C. Schmid was nominated IEEE Fellow, 2012.
  - – Best paper award of Pattern Recognition journal in 2009 for the paper *Description of interest regions with local binary patterns.* M. Heikkila, M. Pietikainen, C. Schmid. Pattern Recognition Volume 42, Issue 3, March 2009, Pages 425-436, http://hal.inria.fr/inria-00548650/en.
  - – We participated in the Multimedia Event Detection track of TrecVid 2012, one of the major benchmarks in automatic video analysis. We ranked 2-nd out of 17 participants for the pre-specified event category task, and first out of 13 participants on the ad-hoc event category task.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Courses taught by team members in 2012:
- Z. Harchaoui, Kernel-based methods for statistical machine learning, 18h, M2, Grenoble University, France.
- Z. Harchaoui, Tutorial on large-scale learning, 1h, ENS-INRIA Visual Recognition and Machine Learning Summer School 2012, Grenoble, France.
- J. Revaud and M. Douze, Multimedia Databases, 18h, M2, ENSIMAG, France.
- C. Schmid, Object recognition and computer vision, 10h, M2, ENS ULM, France.
- C. Schmid and J. Verbeek, Machine Learning & Category Representation, 18h, M2, ENSIMAG, France.
- C. Schmid, Tutorial on image search and classification, 3h, INRIA Visual Recognition and Machine Learning Summer School 2012, Grenoble, France.

9.2.2. Supervision

PhD theses defended in 2012:

- A. Gaidon, *Structured Models for Action Recognition in Real-world Videos - Modèles Structurés pour la Reconnaissance d'Actions dans des Vidéos Réalistes* [1], Université de Grenoble, 25/10/2012, advisers: Z. Harchaoui and C. Schmid.

- T. Mensink, *Apprentissage de Modèles pour la Classification et la Recherche d'Images* [2], Université de Grenoble, 26/10/2012, advisers: J. Verbeek, G. Csurka, and C. Schmid.

- A. Prest, *Weakly supervised methods for learning actions and objects* [3], ETHZ, 4/9/2012, advisers: V. Ferrari, and C. Schmid.

- G. Sharma, *Semantic description of humans in images*, Université de Caen, 17/12/2012, advisers C. Schmid and F. Jurie.

9.2.3. Juries

Participation in PhD defense juries:

- J. Verbeek, jury member of PhD committee for N. Elfiky, Computer Vision Centre, Barcelona, Spain, June 2012.

- C. Schmid, jury member of PhD committee for S. Bak, Université of Sophia-Antipolis, July 2012.

- C. Schmid, jury member of PhD committee for O. Duchenne, ENS Cachan, November 2012.

- C. Schmid, jury member of PhD committee for A. Joulin, ENS Cachan, December 2012.

9.3. Invited presentations

- Z. Akata: Seminar at Computer Vision and Machine Learning group, Institute of Science and Technology, Vienna, Austria, December 2012.

- A. Gaidon: Presentation at MSR-INRIA CVML workshop, Microsoft Research, Cambridge, UK, April, 2012.

- A. Gaidon: Seminar at ETH Zürich, Switzerland, April, 2012.

- A. Gaidon: Seminar at Xerox Research Center Europe (XRCE), Meylan, France, May, 2012.

- Z. Harchaoui: Seminar at Gatsby Neuroscience Unit, UCL, London, March 2012.

- Z. Harchaoui: Presentation at International Symposium in Mathematical Programming, Berlin, August 2012.

- Z. Harchaoui: Seminar at UC Berkeley, September 2012.

- Z. Harchaoui: Presentation at ECML/PKDD Discovery Challenge, Bristol, September 2012.

- Z. Harchaoui: Seminar at Visual Geometry group, Oxford University, October 2012.

- J. Mairal: Seminar at Parietal team, Neurospin, CEA - INRIA, Saclay, France, November 2012.

- J. Mairal: Seminar at Institut Curie, Paris, France, November 2012.

- J. Mairal: Seminar at Willow and Sierra teams, INRIA, Paris, France, November 2012.

- J. Mairal: Seminar at EPFL, Lausanne, Switzerland, November 2012.

- T. Mensink: Seminar at Computer Vision and Machine Learning group, Institute of Science and Technology, Vienna, Austria, March 2012.

- C. Schmid: Workshop on Large Scale Multimedia Search, Los Angeles, January 2012.

- C. Schmid: Seminar at New York University, May 2012.

- C. Schmid: Seminar at Google, Zürich, May 2012.

- C. Schmid: Seminar at ETHZ, Zürich, May 2012.

- C. Schmid: Keynote speaker at ACM International Conference on Multimedia Retrieval (ICMR), Hong Kong, June 2012.

- C. Schmid: Keynote speaker at the International Symposium on Visual Computing, Crete, July 2012.

- C. Schmid: Tutorial on modern features at ECCV 2012, Florence, October 2012.

- C. Schmid: First international workshop on visual analysis and geo-localizaton of large-scale imagery in conjunction with ECCV'12, Florence, October 2012.

- C. Schmid: Keynote speaker at GdR ISIS, Le Touquet, November 2012.

- C. Schmid: Seminar at UC Berkeley, December 2012.

- J. Verbeek: Seminar at TEXMEX group, INRIA, Rennes, France, April 2012.

- J. Verbeek: Seminar at Computer Vision and Machine Learning group, Institute of Science and Technology, Vienna, Austria, June 2012.

- J. Verbeek: Seminar at Computer Vision Centre, Barcelona, Spain, June 2012.

9.4. Popularization

- C. Schmid will present in 2013 the research area of visual recognition to a group of school teachers as well as to a class of high school students.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] A. GAIDON. *Structured Models for Action Recognition in Real-world Videos - Modèles Structurés pour la Reconnaissance d'Actions dans des Vidéos Réalistes*, Université de Grenoble, October 2012.

[2] T. MENSINK. *Apprentissage de Modèles pour la Classification et la Recherche d'Images*, Université de Grenoble, October 2012, http://hal.inria.fr/tel-00752022.

[3] A. PREST. *Weakly supervised methods for learning actions and objects*, Eidgenössische Technische Hochschule Zürich (ETHZ), September 2012, http://hal.inria.fr/tel-00758797.

[4] G. SHARMA. *Semantic description of humans in images*, Université de Caen, December 2012, http://hal.inria.fr/tel-00767699.

Articles in International Peer-Reviewed Journal

[5] M. GUILLAUMIN, T. MENSINK, J. VERBEEK, C. SCHMID. *Face recognition from caption-based supervision*, in "International Journal of Computer Vision", 2012, vol. 96, n⁰ 1, p. 64-82, http://hal.inria.fr/inria-00585834.

[6] H. JÉGOU, F. PERRONNIN, M. DOUZE, J. SÁNCHEZ, P. PÉREZ, C. SCHMID. *Aggregating local image descriptors into compact codes*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", September 2012, http://hal.inria.fr/inria-00633013.

[7] M. MARSZALEK, C. SCHMID. *Accurate Object Recognition with Shape Masks*, in "International Journal of Computer Vision", 2012, vol. 97, n⁰ 2, http://hal.inria.fr/hal-00650941/en.

[8] T. MENSINK, J. VERBEEK, G. CSURKA. *Tree-structured CRF Models for Interactive Image Labeling*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2012, http://hal.inria.fr/hal-00688143.

[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", July 2012, http://hal.inria.fr/hal-00720847.

[10] A. PREST, C. SCHMID, V. FERRARI. *Weakly supervised learning of interactions between humans and objects*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2012, vol. 34, n$^{\text{o}}$ 3, http://hal.inria.fr/inria-00611482/en.

International Conferences with Proceedings

[11] H. BREDIN, J. POIGNANT, M. TAPASWI, G. FORTIER, V. BAC LE, T. NAPOLÉON, G. HUA, C. BARRAS, S. ROSSET, L. BESACIER, J. VERBEEK, G. QUENOT, F. JURIE, E. HAZIM KEMAL. *Fusion of speech, faces and text for person identification in TV broadcast*, in "ECCV Workshop on Information fusion in Computer Vision for Concept Recognition", Florence, Italy, 2012, http://hal.inria.fr/hal-00722884.

[12] R. G. CINBIS, S. SCLAROFF. *Contextual Object Detection using Set-based Classification*, in "European Conference on Computer Vision", Firenze, Italy, October 2012, http://hal.inria.fr/hal-00756638.

[13] R. G. CINBIS, J. VERBEEK, C. SCHMID. *Image categorization using Fisher kernels of non-iid image models*, in "IEEE Conference on Computer Vision & Pattern Recognition", Providence, United States, June 2012, http://hal.inria.fr/hal-00685943.

[14] M. DUDIK, Z. HARCHAOUI, J. MALICK. *Lifted coordinate descent for learning with trace-norm regularization*, in "AISTATS - Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics - 2012", La Palma, Spain, April 2012, http://hal.inria.fr/hal-00756802.

[15] A. GAIDON, Z. HARCHAOUI, C. SCHMID. *Recognizing activities with cluster-trees of tracklets*, in "BMVC", Guildford, United Kingdom, September 2012, http://hal.inria.fr/hal-00722955.

[16] Z. HARCHAOUI, M. DOUZE, M. PAULIN, M. DUDIK, J. MALICK. *Large-scale image classification with trace-norm regularization*, in "IEEE Conference on Computer Vision & Pattern Recognition", Providence, United States, June 2012, p. 3386-3393, http://hal.inria.fr/hal-00728388.

[17] Z. HARCHAOUI, A. JUDITSKY, A. NEMIROVSKI. *Conditional gradient algorithms for machine learning*, in "NIPS Optimization Workshop", December 2012.

[18] T. MENSINK, J. VERBEEK, F. PERRONNIN, G. CSURKA. *Metric Learning for Large Scale Image Classification: Generalizing to New Classes at Near-Zero Cost*, in "ECCV - European Conference on Computer Vision", Florence, Italy, October 2012, http://hal.inria.fr/hal-00722313.

[19] F. PERRONNIN, Z. AKATA, Z. HARCHAOUI, C. SCHMID. *Towards Good Practice in Large-Scale Learning for Image Classification*, in "Computer Vision and Pattern Recognition", Providence, Rhode Island, United States, June 2012, http://hal.inria.fr/hal-00690014.

[20] A. PREST, C. LEISTNER, J. CIVERA, C. SCHMID, V. FERRARI. *Learning Object Class Detectors from Weakly Annotated Video*, in "Computer Vision and Pattern Recognition", Providence, RI, United States, IEEE, June 2012, http://hal.inria.fr/hal-00695940.

[21] J. REVAUD, M. DOUZE, C. SCHMID. *Correlation-Based Burstiness for Logo Retrieval*, in "ACM Multimedia 2012", Nara, Japan, November 2012, http://hal.inria.fr/hal-00728502.

[22] G. SHARMA, F. JURIE, C. SCHMID. *Discriminative Spatial Saliency for Image Classification*, in "Computer Vision and Pattern Recognition", Providence, Rhode Island, United States, June 2012, http://hal.inria.fr/hal-00714311.

[23] G. SHARMA, S. UL HUSSAIN, F. JURIE. *Local Higher-Order Statistics (LHS) for Texture Categorization and Facial Analysis*, in "ECCV - European Conference on Computer Vision", Florence, Italy, August 2012, http://hal.inria.fr/hal-00722819.

Scientific Books (or Scientific Book chapters)

[24] R. BENAVENTE, J. VAN DE WEIJER, M. VANRELL, C. SCHMID, R. BALDRICH, J. VERBEEK, D. LARLUS. *Color Names*, in "Color in Computer Vision", T. GEVERS, A. GIJSENIJ, J. VAN DE WEIJER, J.-M. GEUSEBROEK (editors), Wiley, 2012, http://hal.inria.fr/hal-00640930/en.

[25] Z. HARCHAOUI, F. BACH. *Tree-walk kernels for computer vision*, in "Image Processing and Analysis with Graphs: Theory and Practice", O. LEZORAY, L. GRADY (editors), Digital Imaging and Computer Vision Series, CRC Press, May 2012, http://hal.inria.fr/hal-00756815.

Research Reports

[26] A. GAIDON, Z. HARCHAOUI, C. SCHMID. *Action Detection with Actom Sequence Models*, INRIA, April 2012, n° RR-7930, http://hal.inria.fr/hal-00687312.

[27] T. MENSINK, J. VERBEEK, F. PERRONNIN, G. CSURKA. *Large Scale Metric Learning for Distance-Based Image Classification*, INRIA, September 2012, n° RR-8077, http://hal.inria.fr/hal-00735908.

[28] H. WANG, A. KLÄSER, C. SCHMID, C.-L. LIU. *Dense trajectories and motion boundary descriptors for action recognition*, INRIA, August 2012, n° RR-8050, http://hal.inria.fr/hal-00725627.

Other Publications

[29] S. ARLOT, A. CELISSE, Z. HARCHAOUI. *Kernel change-point detection*, 2012, Available online at arXiv, http://hal.inria.fr/hal-00671174.

[30] M. GUERZHOY, A. HERTZMANN. *Learning Latent Factor Models of Human Travel*, December 2012, NIPS Workshop on Social Network and Social Media Analysis, http://hal.inria.fr/hal-00756192.

[31] D. ONEATA, M. DOUZE, J. REVAUD, S. JOCHEN, D. POTAPOV, H. WANG, Z. HARCHAOUI, J. VERBEEK, C. SCHMID, R. ALY, K. MCGUINESS, S. CHEN, N. O'CONNOR, K. CHATFIELD, O. PARKHI, R. ARANDJELOVIC, A. ZISSERMAN, F. BASURA, T. TUYTELAARS. *AXES at TRECVid 2012: KIS, INS, and MED*, 2013, To appear, http://hal.inria.fr/hal-00746874/en.