IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Activity Report 2011

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 Understanding

Table of contents

Project-Team LEAR

Keywords: Computer Vision, Machine Learning, Recognition, Video

1. Members

Research Scientists

Cordelia Schmid [Team Leader, INRIA Research Director, DR1, HdR]
Zaid Harchaoui [INRIA Researcher, CR2]
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]

Technical Staff

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

PhD Students

Zeynep Akata [Cifre grant Xerox RCE, January '11 – January '14]
Ramazan Cinbis [UJF, INRIA PhD Scholarship, October '10 – October '13]
Florent Dutrech [INPG, MBDA project, September '10 – September '13]
Adrien Gaidon [INPG, Microsoft/INRIA project, October '08 – May '12]
Josip Krapac [University of Caen, ANR project R2I, co-supervision with F. Jurie, January '08 – May '11]
Thomas Mensink [UJF, EU project CLASS Feb. '09 – Sep. '09, Cifre grant Xerox RCE Oct. '09 – Oct. '12]
Dan Oneata [UJF, EU project AXES, QUAERO project, October '11 – October '14]
Danila Potapov [UJF, EU project AXES, QUAERO project, September '11 – August '14]
Alessandro Prest [ETH Zürich, QUAERO project, co-supervision with V. Ferrari, June '09 – May '12]
Gaurav Sharma [University of Caen, ANR project SCARFACE, co-superv. with F. Jurie, Oct. '09 – Oct. '12]

Post-Doctoral Fellow

Jerome Revaud [June '11 – May '12, QUAERO project]

Visiting Scientists

Luca Scarnato [PhD student, Bern University, Switzerland, July '11 – Aug. '11]
Bo Geng [PhD student, Peking University, China, Nov. '11 – April '12]

Administrative Assistants

Anne Pasteur [Secretary INRIA, until August 2011]
Florence Polge [Secretary INRIA, since September 2011]

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

- **Action recognition.** LEAR has developed several successful methods for action recognition [7], [11], [18]. Our approach for action recognition in still images automatically determines objects relevant for an action given a set of training images [7]. In the PASCAL visual object classes challenge 2011 it achieved best results on three out of ten action classes and the best result on average over all classes.

  The approaches [11], [18] model the dynamics of actions in videos. In [18] dense trajectory descriptors are extracted and shown to outperform existing video descriptors. In [11] an "actom sequence model" is introduced, which decomposes actions into sequences of (overlapping) action-units called "actoms". Each actom gathers temporally localized discriminative visual features of the action. This actom sequence model outperformed state-of-the-art approaches on the "Coffee and cigarettes" dataset.
- **Large-scale classification.** LEAR has designed an efficient and scalable approach for large-scale image classification. The approach [10] allows to gracefully scale up to large number of categories and examples while learning the underlying taxonomy of the categories at the same time, by using a

trace-norm regularization penalty. Promising experimental results on subsets of the ImageNet dataset were obtained, where our method outperforms state-of-the-art approaches using 16-Gaussian Fisher vectors. A spatial extension of Fisher vectors [15] allows dimensionality reduction, as does the compression technique presented in [5].

- **INRIA Visual Recognition and Machine Learning Summer School**. This year we co-organized the second edition of the ENS-INRIA Visual Recognition and Machine Learning Summer School in Paris. It attracted a total of 175 participants (31% from France, 50% from Europe and 20% from America and Asia). Next year the summer school will again be organized in Grenoble.

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: Jakob Verbeek [correspondant], Guillaume Fortier.

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, streamlines code developed by different 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). The code will be ported to run in real-time on the mini-pc system of Technosens that implements advanced user interfaces to TV-top videophone systems.

5.2. Large-scale image search

Participants: Matthijs Douze [correspondant], Mohamed Ayari, Cordelia Schmid.

LEAR's image search demonstration was extended to 100M images. The image dataset was provided by Exalead. Search at this scale is possible due to the Fisher vector representation and the pqcodes software. The search time on a single core is about 250 ms.

In collaboration with Hervé Jégou, from the INRIA Texmex team, we stabilized and improved the pqcodes software package. The software was extended to implement matrix multiplications in the PQ-compressed domain. A non-exclusive license on pqcodes was sold to Technicolor. Another agreement is under negotiation with Morpho (a company owned by Safran).

LEAR's implementation of the Fisher descriptor was improved in several ways. A new method to train the GMM was developed and the computation time of second-order derivatives (w.r.t. σ) was significantly reduced. Furthermore, the extraction of dense SIFT descriptors was improved in quality and speed.

5.3. Video descriptors

Participants: Heng Wang, Cordelia Schmid.

We have developed and made on-line available software for video description based on dense trajectories and motion boundary histograms [18]. 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.

6. New Results

6.1. Large-scale image search

6.1.1. Aggregating local image descriptors into compact codes

Participants: Matthijs Douze, Hervé Jégou [INRIA Rennes], Patrick Pérez [Technicolor], Florent Perronnin [Xerox RCE], Jorge Sánchez [Xerox RCE], Cordelia Schmid.

In [5] we consolidate and extend earlier results for large-scale image search. Different ways of aggregating local image descriptors into a vector are compared. The Fisher vector, see Figure 1, is shown to achieve better performance than the reference bag-of-visual words approach for any given vector dimension. Furthermore, we jointly optimize dimensionality reduction and indexing in order to obtain a precise vector comparison as well as a compact representation. The evaluation shows that the image representation can be reduced to a few dozen bytes with good search accuracy. Given such small codes, searching a 100 million image dataset takes about 250 ms on one processor core.

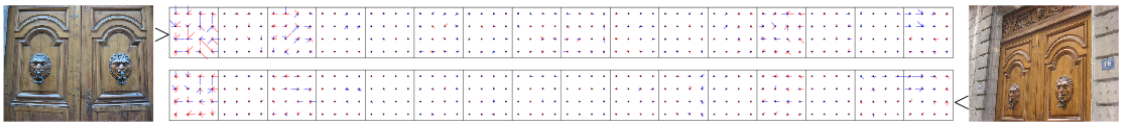

Figure 1. Illustration of the similarity of the Fisher vectors of local image regions despite viewpoint changes.

6.1.2. Searching in one billion vectors: re-rank with source coding

Participants: Laurent Amsaleg [CNRS, IRISA], Matthijs Douze, Hervé Jégou [INRIA Rennes], Romain Tavenard [University Rennes I].

In this work [13] we extend our earlier work [4]. An additional level of processing is added to the product quantizer to refine the estimated distances. It consists in quantizing the difference vector between a point and the corresponding centroid. When combined with an inverted file, this gives three levels of quantization. Experiments performed on SIFT and GIST image descriptors show excellent search accuracy outperforming three state-of-the-art approaches.

6.1.3. Combining attributes and Fisher vectors for efficient image retrieval

Participants: Matthijs Douze, Arnau Ramisa, Cordelia Schmid.

Attributes were recently shown to give excellent results for category recognition. In [9] we demonstrate their performance in the context of image retrieval. We show that combining attributes with Fisher vectors improves performance for retrieval of particular objects as well as categories. Furthermore, we implement an efficient coding technique for compressing the combined descriptor to very small codes. Experimental results show that our approach significantly outperforms the state of the art, even for a very compact representation of 16 bytes per image. We show that attribute features combined with Fisher vectors improve the retrieval of image categories and that those features can supplement text features.

6.1.4. Bag-of-colors for improved image search

Participants: Matthijs Douze, Hervé Jégou [INRIA Rennes], Christian Wengert [Kooaba].

In [19] we investigate the use of color information when used within a state-of-the-art large scale image search system. We introduce a simple color signature generation procedure, used either to produce global or local descriptors. As a global descriptor, it outperforms several state-of-the-art color description methods, in particular the bag-of-words method based on color SIFT. As a local descriptor, our signature is used jointly with SIFT descriptors (no color) to provide complementary information.

6.2. Learning and structuring of visual models

6.2.1. Learning to rank and quadratic assignment

Participants: Thomas Mensink, Jakob Verbeek, Tiberio Caetano [NICTA Canberra].

In [16] we show that the optimization of several ranking-based performance measures, such as precision-at-k and average-precision, is intimately related to the solution of quadratic assignment problems, especially when the score function allows for pairwise label dependencies. Both the task of test-time prediction of the best ranking and the task of constraint generation in estimators based on structured support vector machines can all be seen as special cases of quadratic assignment problems. Although such problems are in general NP-hard, we identify a polynomially-solvable subclass (for both inference and learning) that still enables the modeling of a substantial number of pairwise rank interactions. We show preliminary results on a public benchmark image annotation data set, which indicates that this model can deliver higher performance over ranking models without pairwise rank dependencies. This work was performed during a visit to NICTA Canberra by T. Mensink (March – June, '11) and J. Verbeek (May '11).

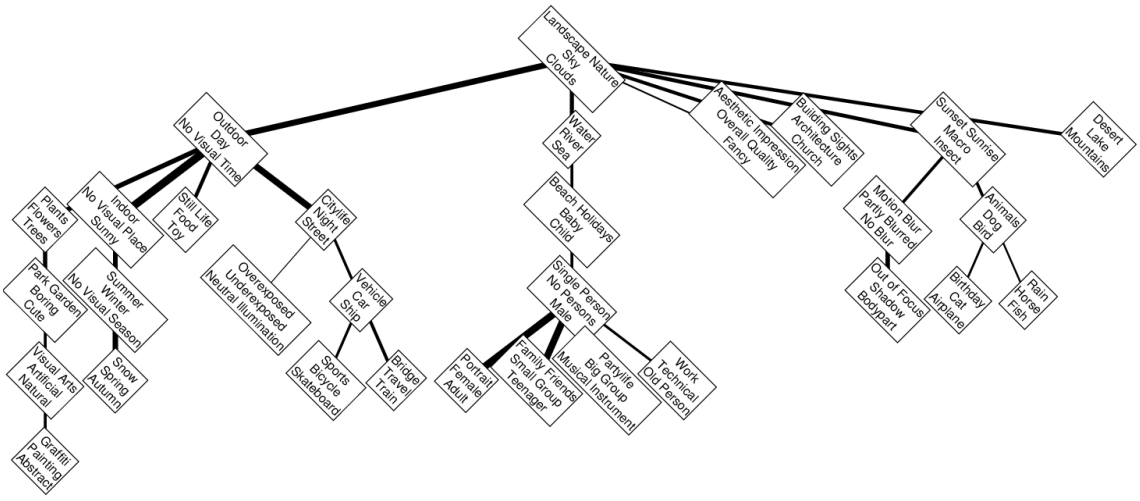

Figure 2. An automatically obtained dependency tree over 96 labels, that contains 3 labels per node.

6.2.2. Learning structured prediction models for interactive image labeling

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

In [25] we propose structured models for image labeling that take into account the dependencies among the image labels explicitly. These models are more expressive than independent label predictors, and lead to more accurate predictions. While the improvement is modest for fully-automatic image labeling, the gain is significant in an interactive scenario where a user provides the value of some of the image labels. Such an interactive scenario offers an interesting trade-off between 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. 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. See Figure 2.

6.2.3. Modeling spatial layout with Fisher vectors for image categorization

Participants: Frédéric Jurie [University of Caen], Josip Krapac, Jakob Verbeek.

In [15] we introduce an extension of bag-of-words image representations to encode spatial layout. Using the Fisher kernel framework we derive a representation that encodes the spatial mean and the variance of image regions associated with visual words. We extend this representation by using a Gaussian mixture model to encode spatial layout, and show that this model is related to a soft-assign version of the spatial pyramid representation. We also combine our representation of spatial layout with the use of Fisher kernels to encode the appearance of local features. Through an extensive experimental evaluation, we show that our representation yields state-of-the-art image categorization results, while being more compact than spatial pyramid representations. In particular, using Fisher kernels to encode both appearance and spatial layout results in an image representation that is computationally efficient, compact, and yields excellent performance while using linear classifiers.

6.2.4. Unsupervised metric learning for face identification in TV video

Participants: Ramazan Cinbis, Jakob Verbeek, Cordelia Schmid.

The goal of face identification is to decide whether two faces depict the same person or not. In [8] we address the identification problem for face-tracks that are automatically collected from uncontrolled TV video data. Face-track identification is an important component in systems that automatically label characters in TV series or movies based on subtitles and/or scripts: it enables effective transfer of the sparse text-based supervision to other faces. We show that, without manually labeling any examples, metric learning can be effectively used to address this problem. This is possible by using pairs of faces within a track as positive examples, while negative training examples can be generated from pairs of face tracks of different people that appear together in a video frame. In this manner we can learn a cast-specific metric, adapted to the people appearing in a particular video, without using any supervision. Identification performance can be further improved using semi-supervised learning where we also include labels for some of the face tracks. We show that our cast-specific metrics not only improve identification, but also recognition and clustering. See Figure 3.

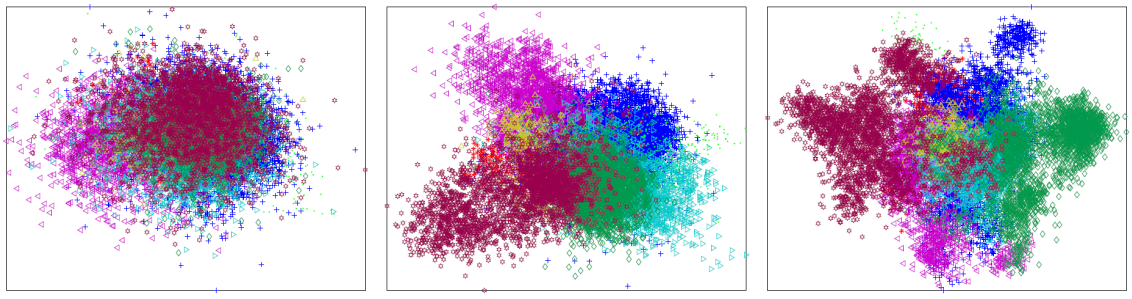

Figure 3. Projections of face signatures projected to two dimensions, using (a) a metric trained on faces detected in still images, (b) using hand labeled faces detected in videos, (c) a metric trained from face tracking results (no manual labeling). Face signatures of different people are color coded. A good face metric can be learned directly from face tracking results, without using any hand labeled examples.

6.2.5. Large-scale image classification

Participants: Miro Dudik [Yahoo! Research], Zaid Harchaoui, Jerome Malick [INRIA Grenoble, BIPOP Team].

We introduced in [10] 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 regular ℓ_1-regularization penalty, we propose a simple and provably efficient coordinate descent algorithm. Furthermore, we showed 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 were obtained, where our approach performed significantly better than state-of-the-art approaches for Fisher vectors with 16 Gaussians.

6.3. Human action recognition

6.3.1. Action recognition by dense trajectories

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

Feature trajectories have shown to be efficient for representing videos. Typically, they are extracted using the KLT tracker or matching SIFT descriptors between frames. However, the quality as well as quantity of these trajectories is often not sufficient. Inspired by the recent success of dense sampling in image classification, in [18] we propose an approach to describe videos by dense trajectories. An overview of our framework is shown in Figure 4. We sample dense points from each frame and track them based on dense optical flow. Our trajectories are robust to fast irregular motions as well as shot boundaries. Additionally, dense trajectories cover the motion information in videos well. We also investigate how to design descriptors to encode the trajectory information. We introduce a novel descriptor based on motion boundary histograms, which is robust to camera motion. This descriptor consistently outperforms other state-of-the-art descriptors, in particular in uncontrolled realistic videos. We evaluate our video description in the context of action classification with a bag-of-features approach. Experimental results show a significant improvement over the state of the art on four datasets of varying difficulty, e.g., KTH, YouTube, Hollywood2 and UCF sports.

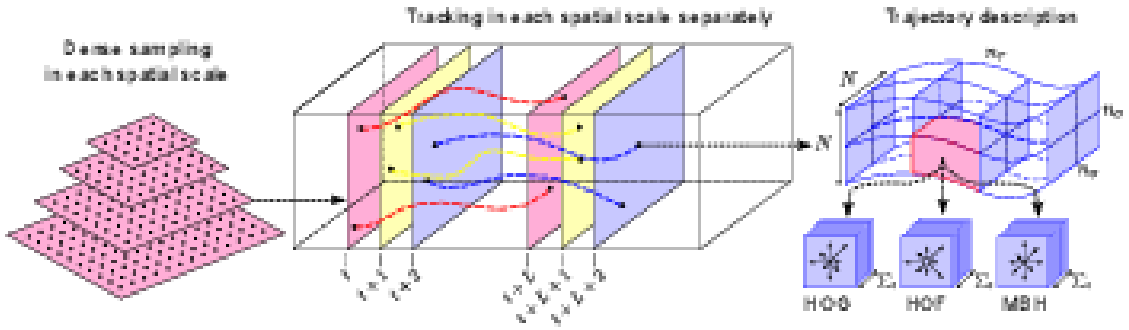

Figure 4. Illustration of dense trajectories extraction and description. Left: dense sampling of feature points at multiple scales; middle: tracking feature points with a dense optical flow field; right: descriptors are computed along the trajectory.

6.3.2. Weakly supervised learning of interactions between humans and objects

Participants: Vittorio Ferrari [ETH Zürich], Alessandro Prest, Cordelia Schmid.

In [7] we introduced a weakly supervised approach for learning human actions modeled as interactions between humans and objects. Our approach is human-centric: we first localize a human in the image and then determine the object relevant for the action and its spatial relation with the human. The model is learned automatically from a set of still images annotated only with the action label. Our approach relies on a human detector to initialize the model learning. For robustness to various degrees of visibility, we build a detector that learns to combine a set of existing part detectors. Starting from humans detected in a set of images depicting the action, our approach determines the action object and its spatial relation to the human. Its final output is a probabilistic model of the human-object interaction, i.e. the spatial relation between the human and the object. We present an extensive experimental evaluation on the sports action dataset from Gupta et al., the PASCAL 2010 action dataset, and a new human-object interaction dataset. In the PASCAL visual object classes challenge 2011 our approach achieved best results on three out of ten action classes and the best result on average over all classes.

6.3.3. Explicit modeling of human-object interactions in realistic videos

Participants: Vittorio Ferrari [ETH Zürich], Alessandro Prest, Cordelia Schmid.

In [26] we introduced an approach for learning human actions as interactions between persons and objects in realistic videos. Previous work typically represents actions with low-level features such as image gradients or optical flow. In contrast, we explicitly localize in space and track over time both the object and the person, and represent an action as the trajectory of the object wrt to the person position. Our approach relies on state-of-the-art approaches for human and object detection as well as tracking. We show that this results in human and object tracks of sufficient quality to model and localize human-object interactions in realistic videos. Our human-object interaction features capture relative trajectory of the object wrt the human. Experimental results on the Coffee & Cigarettes dataset show that (i) our explicit human-object model is an informative cue for action recognition; (ii) it is complementary to traditional low-level descriptors such as 3D-HOG extracted over human tracks. When combining our human-object interaction features with 3D-HOG features, we show to improve over their separate performance as well as over the state of the art. See Figure 5.

6.3.4. Actom sequence models for efficient action detection

Participants: Adrien Gaidon, Zaid Harchaoui, Cordelia Schmid.

In [12] we address the problem of detecting actions, such as drinking or opening a door, in hours of challenging video data. We propose a model based on a sequence of atomic action units, termed "actoms", that are characteristic for the action. Our model represents the temporal structure of actions as a sequence of histograms of actom-anchored visual features. Our representation, which can be seen as a temporally structured extension of the bag-of-features, is flexible, sparse and discriminative. We refer to our model as Actom Sequence Model (ASM). Training requires the annotation of actoms for action clips. 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. We show that our ASM method outperforms the current state of the art in temporal action detection.

6.3.5. A time series kernel for action recognition

Participants: Adrien Gaidon, Zaid Harchaoui, Cordelia Schmid.

In [11] we address the problem of action recognition by describing actions as time series of frames and introduce a new kernel to compare their dynamic aspects. Action recognition in realistic videos has been successfully addressed using kernel methods like SVMs. Most existing approaches average local features over video volumes and compare the resulting vectors using kernels on bags of features. In contrast, we model actions as time series of per-frame representations and propose a kernel specifically tailored for the purpose of action recognition. Our main contributions are the following: (i) we provide a new principled way to compare the dynamics and temporal structure of actions by computing the distance between their auto-correlations,

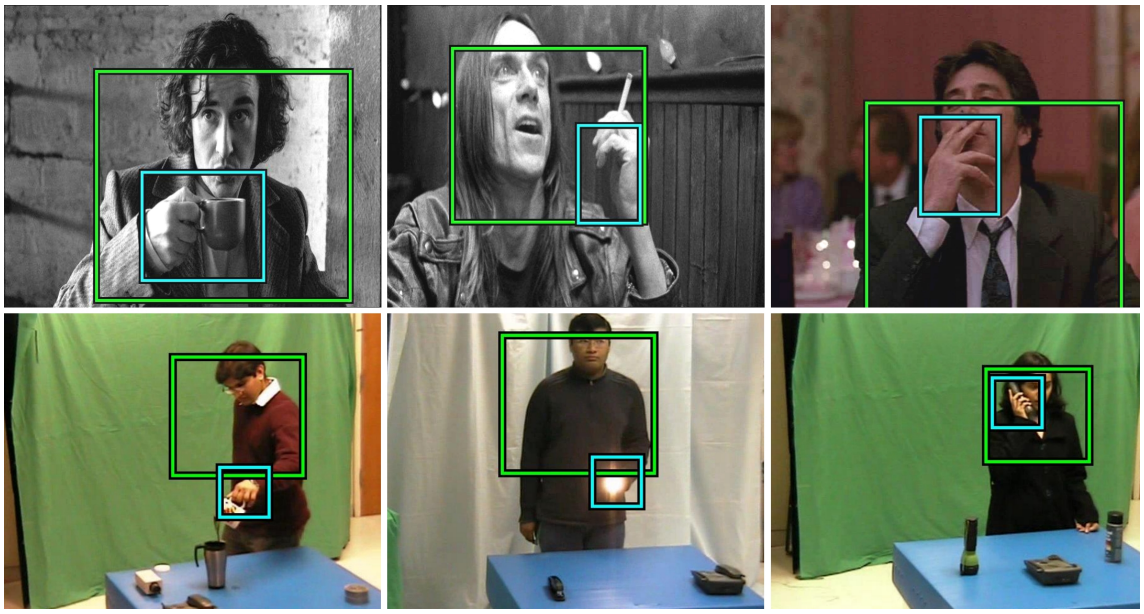

Figure 5. Example results showing the automatically detected human (green) and related object (blue).

(ii) we derive a practical formulation to compute this distance in any feature space deriving from a base kernel between frames, and (iii) we report experimental results on recent action recognition datasets showing that it provides useful complementary information to the average distribution of frames, as used in state-of-the-art models based on bag-of-features.

7. 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: Florent Dutrech, 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. The PhD of Florent Dutrech is funded by this contract.

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

Participants: Adrien Gaidon, Zaid Harchaoui, 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. The PhD student Adrien Gaidon is 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 is 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, implements and evaluates our face recognition system on Technosens hardware. Additional development aims at dealing with hard real-world conditions.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. QUAERO

Participants: Mohamed Ayari, Matthijs Douze, Dan Oneata, Danila Potapov, Alessandro Prest, Cordelia Schmid.

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. Qcompere

Participants: Guillaume Fortier, Cordelia Schmid, Jakob Verbeek.

This three 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 GAIA

Participants: Cordelia Schmid, Jakob Verbeek.

GAIA is an ANR (Agence Nationale de la Recherche) "blanc" project that is running for 4 years starting October 2007. It aims at fostering the interaction between three major domains of computer science—computational geometry, machine learning and computer vision—, for example by studying information distortion measures. The partners are the INRIA project-teams GEOMETRICA and LEAR as well as the University of Antilles-Guyane and Ecole Polytechnique.

8.1.4. ANR Project SCARFACE

Participants: Frédéric Jurie [University of Caen], Cordelia Schmid, Gaurav Sharma.

Video surveillance systems are currently installed in many public areas. As their number increases, the manual analysis becomes impossible. The three-year project SCARFACE (2009-2011) develops tools to automatically access large volumes of video content in order to help investigators solve a crime. These tools will search videos based on human attributes, which describe the suspect. The participant of the project are: the university of Lille the INRIA Imedia team, SpikeNet, EADS, the University of Caen, and LEAR.

8.2. European Initiatives

8.2.1. FP7 European Project AXES

Participants: Ramazan Cinbis, 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.

8.3. International Initiatives

8.3.1. INRIA International Partners

- **NICTA:** In 2010 we initiated a collaboration with the Statistical Machine Learning group at NICTA, Canberra, Australia, i.e., Tiberio Caetano visited LEAR for 4 months. This year PhD student Thomas Mensink spent three months at NICTA, March '11 – June '11, and Jakob Verbeek spent 3 weeks at NICTA in May '11. Results of the collaboration were presented in [16] at the NIPS '11 workshop on Discrete Optimization in Machine Learning.
- **UC Berkeley:** Z. Harchaoui visited UC Berkeley twice in 2011, resp. in January and September 2011. This led to a research collaboration with N. El Karoui on the theoretical analysis of learning algorithms in high-dimensional settings and the influence of the marginal density of the examples on the generalization performance. This collaboration will continue in 2012.

- **ETH Zürich:** We collaborate with V. Ferrari, junior professor at ETH Zürich since his postdoctoral fellowship with the LEAR team in 2006. V. Ferrari and C. Schmid are currently co-supervising a PhD student (A. Prest) on the subject of automatic learning of objects in images and videos [7], [26]. A. Prest is bi-localized between ETH Zürich and INRIA Grenoble.

8.3.2. Visits of International Scientists

8.3.2.1. Internship

- Luca Scarnato, PhD student at Bern University, Switzerland, visited LEAR from July '11 until August '11. He worked on combining color and texture features for image categorization with Jakob Verbeek.
- Bo Geng, PhD student at Peking University, China, is visiting LEAR from November '11 until April '12. He works on attribute-based image retrieval.

9. Dissemination

9.1. Leadership within the scientific community

- Conference, workshop, and summer school organization:
  - C. Schmid: Co-organizer of the INRIA Visual Recognition and Machine Learning Summer School, Paris, July 2011. Attracted a total of 175 participants (31% from France, 50% from Europe and 20% 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.
- Program chair:
  - C. Schmid: ECCV'2012.
- Area chair:
  - J. Verbeek: ECCV'2012.
- Program committees:
  - AISTATS 2011: Z. Harchaoui.
  - CVPR 2011: M. Douze, Z. Harchaoui, C. Schmid, J. Verbeek.
  - ICCV 2011: Z. Harchaoui, J. Verbeek.
  - ICML 2011: Z. Harchaoui.
  - NIPS 2011: Z. Harchaoui.
- Prizes:
  - C. Schmid was nominated IEEE Fellow, 2012.
  - J. Verbeek was awarded an Outstanding Reviewer Award at CVPR '11.

- In the PASCAL visual object classes challenge 2011 our work on human action recognition achieved best results on three out of ten action classes and the best result on average over all classes, see http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2011/ for more details.
- We participated in two tracks of TrecVid 2011, one of the major benchmarks in automatic video analysis. In the Copy Detection task our results were best for 23 of the 56 transformation types. In the Multimedia Event Detection task we ranked 8-th out of 22 participants.

9.2. Teaching

Courses:

- M. Douze and Z. Harchaoui, Multimedia Databases, 3rd year ENSIMAG, 18h.
- C. Schmid, Object recognition and computer vision, Master-2 MVA, ENS ULM, 10h.
- C. Schmid and J. Verbeek, Machine Learning & Category Representation, Master-2 MoSIG, Univ. Grenoble, 18h.
- C. Schmid, Tutorial on image search and classification at the ENS-INRIA Visual Recognition and Machine Learning Summer School, Paris, July 2011.
- C. Schmid, Lecture on local features and large scale search at the 3e cycle romand d'informatique, Geneva, Switzerland, February 2011.
- C. Schmid and Z. Harchaoui, Tutorial on local features, image search and classification as well as learning at the Winter Research School at ENS Lyon, January 2011.

PhD theses:

- J. Krapac, *Représentations d'images pour la recherche et la classification d'images*, Université de Caen, 11/7/2011, advisors: F. Jurie and J. Verbeek.
- H. Harzallah, *Contribution à la localisation et à la reconnaissance d'objets dans les images*, Université de Grenoble, 16/9/2011, advisors: F. Jurie and C. Schmid.

9.3. Invited presentations

- Z. Akata: Presentation at Hands-on Image Processing (HOIP'11) at Parque Tecnológico de Vizcaya, November 2011.
- G. Fortier: Technical demonstration at Forum 4i (Innovation, Industrie, Investissement, International), Grenoble, May 2011.
- G. Fortier: Technical demonstration at Grenoble Innovation Fair, October 2011.
- G. Fortier: Technical demonstration at Rencontres INRIA-Industrie, Rennes, November 2011.
- A. Gaidon: Technical demonstration at the Microsoft Research / INRIA forum, Paris, April 2011.
- A. Gaidon: Seminar at the Computer Vision Center, Autonomous University of Barcelona, Spain, May 2011.
- A. Gaidon: Presentation at the "Journée perception de l'homme et reconnaissance d'actions", GdR ISIS, CNRS, Paris, June 2011.
- Z. Harchaoui: Seminar at University of California Berkeley, January 2011.
- Z. Harchaoui: Presentation at Stat'Learn, Grenoble, March 2011.
- Z. Harchaoui: Presentation at GDR Isis, Paris, April 2011.
- Z. Harchaoui: Seminar at University Paris VI, May 2011.
- Z. Harchaoui: Seminar at Xerox Research Center Europe, September, Meylan, 2011.
- Z. Harchaoui: Seminar at Kyoto University, Japan, November 2011.

- Z. Harchaoui: Presentation at INRIA Workshop on Statistical Learning, Paris, December 2011.
- Z. Harchaoui: Presentation at NIPS Workshop, Sierra Nevada, Spain, December 2011.
- T. Mensink: Seminar at Statistical Machine Learning group, NICTA Canberra, Australia, March 2011.
- T. Mensink: Seminar at Calvin group, ETH Zürich, Switzerland, September 2011.
- C. Schmid: Presentation at Colloquium J. Morgenstern, Sophia-Antipolis, December 2011.
- C. Schmid: Presentation at NIPS Workshop, Granada, Spain, December 2011.
- C. Schmid: Presentation at Symposium on Applied Perception in Graphics and Visualization, Toulouse, August 2011.
- C. Schmid: Presentation at Frontiers in Computer Vision Workshop, MIT, August 2011.
- J. Verbeek: Presentation at 2nd IST Workshop on Computer Vision and Machine Learning, Institute of Science and Technology, Vienna, Austria, October 2011.
- J. Verbeek: Presentation at Workshop on 3D and 2D Face Analysis and Recognition, Ecole Centrale de Lyon, January 2011.
- J. Verbeek: Seminar at Statistical Machine Learning group, NICTA, Canberra, Australia, May 2011.
- J. Verbeek: Seminar at Canon Information Systems Research Australia, Sydney, Australia, May 2011.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] H. HARZALLAH. *Contribution à la localisation et à la reconnaissance d'objets dans les images*, Université de Grenoble, September 2011, http://hal.inria.fr/tel-00623278/en.

[2] J. KRAPAC. *Représentations d'images pour la recherche et la classification d'images*, Université de Caen, July 2011, http://hal.inria.fr/tel-00650998/en.

Articles in International Peer-Reviewed Journal

[3] M. GUILLAUMIN, T. MENSINK, J. VERBEEK, C. SCHMID. *Face recognition from caption-based supervision*, in "International Journal of Computer Vision", 2011, To appear, http://hal.inria.fr/inria-00585834/en.

[4] H. JÉGOU, M. DOUZE, C. SCHMID. *Product Quantization for Nearest Neighbor Search*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", January 2011, vol. 33, n^{o} 1, p. 117–128 [*DOI :* 10.1109/TPAMI.2010.57], http://hal.inria.fr/inria-00514462/en.

[5] 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", 2011, To appear, http://hal.inria.fr/inria-00633013/en.

[6] M. MARSZALEK, C. SCHMID. *Accurate Object Recognition with Shape Masks*, in "International Journal of Computer Vision", 2011, To appear, http://hal.inria.fr/hal-00650941/en.

[7] A. PREST, C. SCHMID, V. FERRARI. *Weakly supervised learning of interactions between humans and objects*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2011, To appear, http://hal.inria.fr/inria-00611482/en.

International Conferences with Proceedings

[8] R. G. CINBIS, J. VERBEEK, C. SCHMID. *Unsupervised Metric Learning for Face Identification in TV Video*, in "International Conference on Computer Vision", Barcelona, Spain, November 2011, http://hal.inria.fr/inria-00611682/en.

[9] M. DOUZE, A. RAMISA, C. SCHMID. *Combining attributes and Fisher vectors for efficient image retrieval*, in "IEEE Conference on Computer Vision & Pattern Recognition", Colorado Springs, United States, June 2011, http://hal.inria.fr/inria-00566293/en.

[10] M. DUDIK, Z. HARCHAOUI, J. MALICK. *Coordinate descent for learning with trace-norm regularization* , in "AISTATS", 2012.

[11] A. GAIDON, Z. HARCHAOUI, C. SCHMID. *A time series kernel for action recognition*, in "British Machine Vision Conference", Dundee, United Kingdom, August 2011, http://hal.inria.fr/inria-00613089/en.

[12] A. GAIDON, Z. HARCHAOUI, C. SCHMID. *Actom Sequence Models for Efficient Action Detection*, in "IEEE Conference on Computer Vision & Pattern Recognition", Colorado Springs, United States, June 2011, http://hal.inria.fr/inria-00575217/en.

[13] H. JÉGOU, R. TAVENARD, M. DOUZE, L. AMSALEG. *Searching in one billion vectors: re-rank with source coding*, in "International Conference on Acoustics, Speech and Signal Processing", Prague, Czech Republic, May 2011, http://hal.inria.fr/inria-00566883/en.

[14] J. KRAPAC, J. VERBEEK, F. JURIE. *Learning Tree-structured Descriptor Quantizers for Image Categorization*, in "British Machine Vision Conference", Dundee, United Kingdom, September 2011, http://hal.inria.fr/inria-00613118/en.

[15] J. KRAPAC, J. VERBEEK, F. JURIE. *Modeling Spatial Layout with Fisher Vectors for Image Categorization*, in "International Conference on Computer Vision", Barcelona, Spain, November 2011, http://hal.inria.fr/inria-00612277/en.

[16] T. MENSINK, J. VERBEEK, T. CAETANO. *Learning to Rank and Quadratic Assignment*, in "NIPS Workshop on Discrete Optimization in Machine Learning", Granada, Spain, December 2011, http://hal.inria.fr/hal-00645623/en.

[17] T. MENSINK, J. VERBEEK, G. CSURKA. *Learning structured prediction models for interactive image labeling*, in "IEEE Conference on Computer Vision & Pattern Recognition (CVPR '11)", Colorado Springs, United States, June 2011, http://hal.inria.fr/inria-00567374/en.

[18] H. WANG, A. KLÄSER, C. SCHMID, L. CHENG-LIN. *Action Recognition by Dense Trajectories*, in "IEEE Conference on Computer Vision & Pattern Recognition", Colorado Springs, United States, June 2011, p. 3169-3176, http://hal.inria.fr/inria-00583818/en.

[19] C. WENGERT, M. DOUZE, H. JÉGOU. *Bag-of-colors for improved image search*, in "ACM Multimedia", Scottsdale, United States, October 2011, http://hal.inria.fr/inria-00614523/en.

Conferences without Proceedings

[20] M. AYARI, J. DELHUMEAU, M. DOUZE, H. JÉGOU, D. POTAPOV, J. REVAUD, C. SCHMID, J. YUAN. *INRIA@TRECVID'2011: Copy Detection & Multimedia Event Detection*, in "TRECVID", Gaithersburg, United States, December 2011, http://hal.inria.fr/hal-00648016/en.

Scientific Books (or Scientific Book chapters)

[21] 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, 2011, To appear, http://hal.inria.fr/hal-00640930/en.

[22] J. WEIJER, T. GEVERS, C. SCHMID, A. GIJSENIJ. *Color Ratios*, in "Color in Computer Vision", T. GEVERS, A. GIJSENIJ, J. VAN DE WEIJER, J.-M. GEUSEBROEK (editors), Wiley, 2011, http://hal.inria.fr/hal-00650946/en.

Research Reports

[23] H. JÉGOU, M. DOUZE, C. SCHMID. *Exploiting descriptor distances for precise image search*, INRIA, June 2011, http://hal.inria.fr/inria-00602325/en.

[24] J. KRAPAC, J. VERBEEK, F. JURIE. *Spatial Fisher Vectors for Image Categorization*, INRIA, August 2011, n^{o} RR-7680, http://hal.inria.fr/inria-00613572/en.

[25] T. MENSINK, J. VERBEEK, G. CSURKA. *Weighted Transmedia Relevance Feedback for Image Retrieval and Auto-annotation*, INRIA, December 2011, n^{o} RT-0415, http://hal.inria.fr/hal-00645608/en.

[26] A. PREST, V. FERRARI, C. SCHMID. *Explicit modeling of human-object interactions in realistic videos*, INRIA, September 2011, n^{o} RT-0411, http://hal.inria.fr/inria-00626929/en.

[27] O. YAKHNENKO, J. VERBEEK, C. SCHMID. *Region-Based Image Classification with a Latent SVM Model*, INRIA, July 2011, n^{o} RR-7665, http://hal.inria.fr/inria-00605344/en.