INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team LEAR

Learning and recognition in vision

Grenoble - Rhône-Alpes

Theme : Vision, Perception and Multimedia Understanding

Table of contents

LEAR is a joint team of INRIA and the LJK laboratory, a joint research unit of the Centre National de Recherche Scientifique (CNRS), the Institut National Polytechnique de Grenoble (INPG) and the Université Joseph Fourier (UJF).

1. Team

Research Scientists

Cordelia Schmid [Team Leader, INRIA Research Director, DR1, HdR]
Zaid Harchaoui [INRIA Researcher, CR2, since January '10]
Rémi Ronfard [INRIA Researcher, CR1, HdR]
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 '11, QUAERO project]
Matthijs Douze [Since January'05, QUAERO project since '08, permanent engineer SED since November '10]
Guillaume Fortier [October '10 – April '12, ITI Visages project]

PhD Students

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 – September '11]
Matthieu Guillaumin [INPG, Ministry of research grant, September '07 – September '10]
Hedi Harzallah [INPG, MBDA project, August '07 – July '10]
Alexander Kläser [INPG, EU project CLASS, November '06 – July '10]
Josip Krapac [University of Caen, ANR project R2I, co-supervision with F. Jurie, Jan. '08 – March '10]
Jörg Liebelt [INPG, EADS scholarship, co-superv. with R. Westermann, TU Munich, Oct. '06 – Oct. '10]
Thomas Mensink [UJF, EU project CLASS Feb. '09 – Sep. '09, Cifre grant Xerox RCE Oct. '09 – Oct. '12]
Alessandro Prest [ETH Zürich, QUAERO project, co-supervision with V. Ferrari, Jun. '09 – Jun. '12]
Gaurav Sharma [University of Caen, ANR project SCARFACE, co-superv. with F. Jurie, Oct. '09 – Oct. '12]

Post-Doctoral Fellows

Arnau Ramisa [IIIA-CSIC September '09 – December '09, QUAERO project, January '10 – December '10]
Oksana Yakhnenko [INRIA, ANR project GAIA, November '09 – November '10]

Visiting Scientists

Tiberio Caetano [Senior Researcher, Statistical Machine Learning Group NICTA, Sep. '10 – Dec. '10]
Miroslav Dudík [Postdoctoral Researcher, Machine Learning Department CMU, Sep.'10 – Nov.'10]
Heng Wang [PhD student, LIAMA, Inst. of Automation, Chinese Ac. of Science, Mar.'10 – Nov.'10]

Administrative Assistant

Anne Pasteur [Secretary INRIA]

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. Ideally, methods will use this data to automatically learn actions despite the noisy labels.

2.2. Highlights

- **Excellent results in ImageCLEF evaluation campaign**. LEAR participated together with Xerox RCE in the Photo Annotation task of the ImageCLEF 2010 evaluation campaign. Our results were ranked first among 16 international participating research teams from industry and academia. For more information see http://www.imageclef.org/2010/PhotoAnnotation.

- **Human action recognition**. In the PASCAL VOC 2010 our work on human action recognition achieved best results on three out of nine action classes, see http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/ for more details. In the ECCV'10 International Workshop on Sign, Gesture, and Activity, our paper [20] was awarded the best paper prize.
- **Visual Recognition and Machine Learning Summer School.** The first INRIA Visual Recognition and Machine Learning Summer School took place at our institute in Grenoble, from July 26 to July 30. The school had 150 international attendees, and was organized by the LEAR team and the WILLOW team at INRIA Rocquencourt. For more information see http://www.di.ens.fr/willow/events/cvml2010/.

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 interest points.

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). The ANR R2I investigates how to search conjointly on images and text. In a collaboration with Xerox Research Centre 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 upcoming 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

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

the so-called user-generated content sites such as Flickr, YouTube or DailyMotion. Another type of application is the detection of specific content from images and videos, which can be used for a large number of problems, such as copyright enforcement for videos, or given an image of a movie poster finding relevant information such as when and where the movie is playing. Transfer of such techniques is the goal of the creation of the start-up MilPix, to which our current technologies for image search are licenced. 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 in infrared images observed from airplanes.

5. Software

5.1. Large-scale image search

Our large-scale image indexing software was extended this year by adding a new image description which aggregates local descriptors [19]. This description is shown to be efficiently indexed with the product quantizer [29]. The product quantizer was registered at the "Agence pour la protection des programmes" (APP) under IDDN.FR.001.220012.000.S.P.2010.000.10000.

Based on this approach [19], we developed an image search demonstrator that is able to index 10 million images on a laptop computer. An image is represented by only 21 bytes and image search in 10 million images takes about 20ms. The images themselves are stored on a large-capacity external hard disk. The indexing structure returns a shortlist of 300 images which are retrieved from the hard disk together with their local descriptors. The shortlist is reordered using this information based on geometric verification. The demonstrator obtained the best demo award at RFIA 2010. It also was shown at the CVPR'2010 demonstration session and at the Quaero meeting in Rennes in April 2010.

5.2. Face recognition

In a collaboration with Technosens (a start-up based in Grenoble) we are developing an efficient face recognition library. During 18 months an engineer, financed by INRIA's technology transfer program, will streamline 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.

5.3. Datasets

Relevant datasets are important to assess recognition methods. In addition to the datasets we created previously, we released several new datasets this year together with the pre-processed image descriptors. Our publicly accessible datasets are available at http://lear.inrialpes.fr/data.

Labeled Yahoo! news. This data set extends the Labeled Faces in the Wild data set that was introduced in 2008 for evaluation of face verification in uncontrolled settings, and has become the de-facto standard since. Our extended data set contains 31147 detected faces of 5873 different people in 20071 images downloaded from Yahoo!News. Its annotation provides all associations of faces in the images with names in the captions (detected using face detectors, and named-entity detectors respectively). The data set can be used for more thorough evaluation of face verification methods (that determine for two faces whether they are of the same person or not), but more importantly it can be used for evaluation of methods that automatically link names and faces detected in the image and in the caption. This data set was used in our papers [17], [31].

Labeled web queries data set. This data set contains 71478 images and text meta-data in XML format retrieved for 353 text queries using a commercial web image search system. The annotation contains relevance labels for each image indicating whether the image is relevant to the text query for which it was retrieved or not. This data set was used in our paper [21] to evaluate a new method to filter search engine results using the image content, which is generally ignored in today's search engines.

6. New Results

6.1. Large-scale image search

6.1.1. Aggregating local descriptors into a compact image representation

Participants: Matthijs Douze, Hervé Jégou [INRIA Rennes], Patrick Pérez [Technicolor], Cordelia Schmid.

We address the problem of image search on a very large scale, where three constraints have to be considered jointly: the accuracy of the search, its efficiency, and the memory usage of the representation. We proposed in [19] an efficient way of aggregating local image descriptors into a vector of limited dimension. We then show how to jointly optimize the dimension reduction and the indexing algorithm, so that it best preserves the quality of vector comparison. The evaluation shows that our approach significantly outperforms the state of the art: the search accuracy is comparable to the bag-of-features approach for an image representation that fits in 20 bytes. Search in a 10 million image dataset takes about 20ms.

6.1.2. Compact video description with precise temporal alignment

Participants: Matthijs Douze, Hervé Jégou [INRIA Rennes], Patrick Pérez [Technicolor], Cordelia Schmid.

In [15] we introduce a very compact yet discriminative video description, which allows example-based search in a large number of frames corresponding to thousands of hours of video. Frame descriptors are encoded using a time-aware hierarchical indexing structure. A modified temporal Hough voting scheme is used to rank the retrieved database videos and find segments that match the query. Experimental results on the TRECVID 2008 copy detection videos and a set of 38000 videos from YouTube show that our method offers an excellent trade-off between search accuracy, efficiency and memory usage.

6.1.3. Product quantization for nearest neighbor search

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

In [29] we introduce an approach based on product quantization for approximate nearest neighbor search. The space is decomposed into a Cartesian product of low dimensional subspaces and each subspace is quantized separately. A vector is represented by a short code composed of its subspace quantization indices. The Euclidean distance between two vectors can be computed in the compressed domain. An asymmetric version increases precision, as it computes the approximate distance between a vector and a code.

6.2. Learning and structuring of visual models

6.2.1. Improving web image search results using query-relative classifiers

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

Previous image re-ranking methods which take into account visual features require separate training for every new query, and are therefore unsuitable for real-world web search applications. Our approach [21] instead learns a single generic classifier, based on 'query-relative' features. The features combine textual information about the occurrence of the query terms and other words found to be related to the query, and visual information derived from a visual histogram image representation. We can train the model once, using whatever annotated data is available, then use it to make predictions for previously unseen test classes.

6.2.2. Semi-supervised image categorization

Participants: Matthieu Guillaumin, Cordelia Schmid, Jakob Verbeek.

In [16] we study the problem of image categorization using semi-supervised techniques. We consider a scenario where keywords are associated with the training images, e.g. as found on photo sharing websites. We learn classifiers for images alone, but use keywords associated with labeled and unlabeled training images to improve the classifier using semi-supervised learning. In our experiments we consider 58 categories from the PASCAL VOC'07 and MIR Flickr sets. For most categories we find our semi-supervised approach to perform better than an approach that uses only labeled images. Figure 1 shows example images with their keywords and category labels as used in our experiments.

Tags: desert, nature, landscape, sky
Labels: clouds, plant life, sky, tree

Tags: rose, pin

Tags: india
Labels: flower, *Labels*: cow

Tags: aviation, airplane, airport
Labels: aeroplane

Figure 1. Example images used in our experiments, together with the user tags and category labels.

6.2.3. Trans-media relevance feedback for image auto-annotation

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

Many image auto annotation methods are based on visual similarity between images to be annotated and images in a training corpus, i.e. by transferring the annotations of the most similar training images. In [24] we consider using also similarities among the training images, both visual and textual, to derive pseudo relevance models, as well as trans-media relevance models. On two widely used datasets (COREL and IAPR) we show experimentally that the pseudo-relevance models improve the annotation accuracy.

6.2.4. Multiple instance metric learning from faces labeled by image captions

Participants: Matthieu Guillaumin, Cordelia Schmid, Jakob Verbeek.

In [17] we proposed a method to learn metrics for face verification (to determine if two faces are of the same person or not). Our method does not require the training images to be labeled by identity, but can use a set of names corresponding to a set of faces observed in the image. To evaluate our approach, we introduced a large and challenging data set, Labeled Yahoo! News, see Section 5.3. In [7], [31] we have used this data set to evaluate methods that automatically infer which name belongs to which face, demonstrating the importance of metrics learned for face verification to solve this problem.

6.2.5. Face recognition from caption-based supervision

Participants: Matthieu Guillaumin, Thomas Mensink, Cordelia Schmid, Jakob Verbeek.

In a forthcoming journal paper [7], [31] the goal is to take a collection of images with captions, and to associate names found in the caption with faces found in the image. Both name and face detection are performed automatically, and the association problem can be seen as a form of semi-supervised learning. Through extensive experiments we evaluate various techniques to solve for the associations, and the impact of metric learning in this context. Metric learning is used to find a distance metric between faces that is robust to changes in expression and pose, while being sensitive to subtle appearance variations from one person to another. We find that using metric learning leads to important improvements in recognition performance, for all name-face association techniques that were evaluated.

6.2.6. Automatic image annotation

Participants: Gabriela Csurka [XRCE], Matthieu Guillaumin, Thomas Mensink, Florent Perronnin [XRCE], Jorge Sánchez [XRCE], Cordelia Schmid, Jakob Verbeek.

In [13] we evaluate different variants of our image annotation model TagProp (developed in 2009, and presented this year at RFIA 2010 [28]) with experiments on the MIR Flickr set, and compare with an approach that learns a separate SVM classifier for each annotation term. We also consider using Flickr tags to train our models, both as additional features and as training labels. See Figure 2 for several example images with automatic annotations. In [23] we describe our submission to the ImageCLEF Visual Concept Detection and Annotation Task, where we combined our individual state-of-the-art approaches: the Fisher vector image representation with the TagProp method for image auto-annotation.

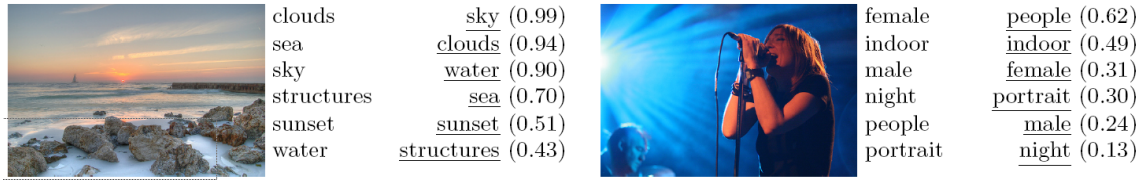

Figure 2. For each image we show the manual annotation (left) and the automatically predicted one (right), where we give the confidence value for each predicted word, and underline it when it appears in the manual annotation.

6.2.7. A 3D geometric model for multi-view object class detection

Participants: Jörg Liebelt, Cordelia Schmid.

We developed a new approach for multi-view object class detection [22] that discriminatively learns the object appearance from real images, and learns the 3D geometry generatively from synthetic models. In contrast to other methods, neither tedious manual part annotation of training images nor explicit appearance matching between synthetic and real training data is required, which results in high geometric fidelity and in increased flexibility. On current state-of-the-art benchmarks our approach outperforms previously published results for viewpoint estimation.

6.3. Human action recognition

6.3.1. Survey of methods for action representation, segmentation and recognition

Participants: Edmond Boyer [Perception/INRIA], Rémi Ronfard, Daniel Weinland [Deutsche Telekom].

In [34], [12] we classify approaches to action recognition in over 150 papers with respect to how they represent the spatial and temporal structure of actions, segment and recognize actions from a continuous video stream, and handle variations in camera viewpoint. Important issues that must be addressed in future work are scalability of action recognition systems with respect to vocabulary size; recognition in the presence of unknown actions; scenes containing multiple persons; and interactions between people.

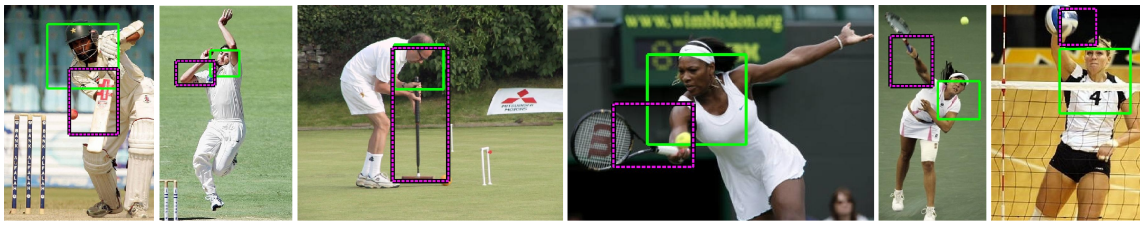

Figure 3. Example results showing the automatically detected human (green) and object (pink) for actions cricket batting, cricket bowling, croquet, tennis forehand, tennis serve, and volleyball.

6.3.2. Weakly supervised learning of interactions between humans and objects

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

We introduced a weakly supervised approach for learning human actions modeled as interactions between humans and objects [33]. 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. See Figure 3 for example human-object detections.

6.3.3. Human focused action localization in video

Participants: Alexander Kläser, Marcin Marszałek [University of Oxford], Cordelia Schmid, Andrew Zisserman [University of Oxford].

We developed a method [20] to localize human actions in uncontrolled video, such as Hollywood movies. We first extract spatio-temporal human tracks and then detect actions within these using a sliding window classifier. To localize actions within the extracted tracks, we introduce a spatio-temporal 3D histogram-of-gradient based descriptor adapted to the track. We obtain significant improvements over the current state of the art. Figure 4 shows the top three drinking detections in the movie *Coffee and Cigarettes*.

Figure 4. Two top drinking detections within the movie Coffee and Cigarettes, the green rectangle indicates the spatial extent of the detection.

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. Hervé Jégou and Cordelia Schmid are the scientific advisers of MILPIX.

7.2. MDBA Aerospatiale

Participants: Florent Dutrech, Hedi Harzallah, 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 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, LEAR, and FLUMINANCE 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 Centre Europe

Participants: Thomas Mensink, 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 in databases that contain documents in different modalities, such as texts, images, or videos, and documents that contain a combination of these. Given a query in one or multiple media, the goal is to retrieve documents in other media. In addition to retrieval we also consider visualization, clustering, and classification of documents in such databases. 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 will start in early 2011.

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, will implement and evaluate our face recognition system on Technosens hardware. Additional development will be aimed at dealing with hard real-world conditions.

8. Other Grants and Activities

8.1. National Projects

8.1.1. QUAERO

Participants: Mohamed Ayari, Matthijs Douze, Alessandro Prest, Arnau Ramisa, 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: 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, Oksana Yakhnenko.

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 R2I

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

R2I (Recherche d'Image Interactive) is an ANR "masse de données et connaissances" project that is running for 3 years starting in January 2008. R2I aims at designing methods for interactive image search, i.e., to extract semantics from images, to cluster similar images and to enable user interaction via semantic concepts related to images. The partners are the company Exalead, a leader in the area of corporate network indexing and a specialist for user-centered approaches, the INRIA project-team Imedia, a research group with a strong background in interactive search of multi-media documents, as well as LEAR and the University of Caen. See https://r2i.greyc.fr/ for more details.

8.1.5. 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. International Projects

8.2.1. FP7 European Project AXES

Participants: Ramazan Cinbis, Cordelia Schmid, Jakob Verbeek.

This 4-year project will start 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.

8.2.2. FP7 European Network of Excellence PASCAL 2

Participants: Adrien Gaidon, Matthieu Guillaumin, Zaid Harchaoui, Cordelia Schmid, Jakob Verbeek.

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

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, Grenoble, July 2010.
  - J. Verbeek: Co-organizer of NIPS'10 Workshop on Machine Learning for Next Generation Computer Vision Challenges, Whistler BC Canada, December 2010.
- Editorial boards:
  - C. Schmid: International Journal of Computer Vision, since 2004.
  - C. Schmid: Foundations and Trends in Computer Graphics and Vision, since 2005.
- Program chair:
  - C. Schmid: ECCV'2012.
- Area chair:
  - C. Schmid: CVPR 2010.
  - C. Schmid: ECCV 2010.
  - C. Schmid: RFIA 2010.
- Program committees:
  - 3DPVT 2010: R. Ronfard.
  - ACM MM 2010: M. Douze.
  - CVPR 2010: M. Douze, Z. Harchaoui, C. Schmid, J. Verbeek.
  - ECCV 2010: M. Douze, J. Verbeek.
  - ICIP 2010: R. Ronfard.
  - ICML 2010: Z. Harchaoui.
  - ICRA 2010: A. Ramisa.
  - IROS 2010: A. Ramisa.
  - NIPS 2010: Z. Harchaoui, F. Jurie, J. Verbeek.
  - RFIA 2010: J. Verbeek.
- Prizes:
  - Our submissions to the ImageCLEF evaluation campaign for the "Photo Annotation" track obtained a first place among 16 participants, see http://imageclef.org/2010 and [23].

  - In the PASCAL visual object classes challenge 2010 our work on human action recognition achieved best results on three out of nine action classes, see http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/ for more details.
  - In the ECCV'10 International Workshop on Sign, Gesture, and Activity, our paper [20] was awarded the best paper prize.
  - At RFIA 2010 we received the best demonstration award for the demonstration "10 million images on my laptop" [29].

- Other:
  - C. Schmid and J. Verbeek were member of recruiting committees for ENSIMAG, 2010.
  - C. Schmid is a member of INRIA's "Commission d'Évaluation". She was in charge of the CR2/CR1 2010 recruiting committees at INRIA Grenoble, Rhône-Alpes.
  - C. Schmid is a member of the "conseil de l'agence d'évaluation de la recherche et de l'enseignement supérieur (AERES)" since March 2007.
  - C. Schmid is a member of the INRIA Grenoble, Rhône-Alpes local scientific committee (bureau du comité des projets) since 2007.

9.2. Teaching

- M. Douze and Z. Harchaoui, Multimedia Databases, 3rd year ENSIMAG, 18h.
- M. Guillaumin, several exercise classes in the context of a "monitorat" at ENSIMAG, in total 64h.
- Z. Harchaoui, C. Schmid, and J. Verbeek, Tutorials at the INRIA Visual Recognition and Machine Learning Summer School, Grenoble, July 2010.
- 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.

9.3. Invited presentations

- M. Douze, INRIA-EADS meeting, Paris, April 2010.
- M. Douze, MPEG workshop on image indexing, Geneva, July 2010.
- Z. Harchaoui, *Learning distinguishing marks for classification*, seminar Willow team, INRIA & ENS, Paris, January 2010.
- Z. Harchaoui, *Temporal segmentation with Kernel change-point detection*, seminar CBLL/VLG group, NYU, New York, February 2010.
- Z. Harchaoui, *Change-point detection with kernels*, seminar Willow team, INRIA & ENS, Paris, March 2010.
- Z. Harchaoui, *Kernel Change-point Analysis*, Seminar fur Statistik, ETH, Zürich, December 2010.
- R. Ronfard, *Machine Learning Methods in Visual Recognition of Gesture and Action*, Spring school of the 3eme cycle romand in Computer Science, Gesture Recognition, University of Fribourg, Switzerland, June 2010.
- C. Schmid, CVPR area chair meeting workshop, University of Maryland, February 2010.
- C. Schmid, seminar at Centre de Mathématiques et de Leurs Applications, Ecole Normale Supérieure de Cachan, Paris, April 2010.
- C. Schmid, ECCV area chair colloquium, Paris, June 2010.
- C. Schmid, seminar at Oxford University, July 2010.

- C. Schmid, seminar at New York University, July 2010.
- C. Schmid, journée d'échanges et de formation, LERTI, INRIA, Grenoble,September 2010.
- C. Schmid, keynote speaker at Coresa 2010, Lyon, October 2010.
- J. Verbeek, *TagProp: a discriminatively trained nearest neighbor model for image auto-annotation*, Seminar at University of Oxford, February 2010.
- J. Verbeek, *Metric learning approaches for caption-based face recognition in uncontrolled settings*, ECCV Workshop on Face Detection: Where are we, and what next?, Hersonissos, Greece, September 2010.
- J. Verbeek, *Metric learning approaches for image annotation and face verification*, Seminar Laboratoire TIMC-IMAG, Learning: Models and Algorithms team, Grenoble, October 2010.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] M. GUILLAUMIN. *Exploiting multimodal data for image understanding*, Université de Grenoble, September 2010, http://lear.inrialpes.fr/pubs/2010/Gui10.

[2] A. KLÄSER. *Learning human actions in video*, Université de Grenoble, July 2010, http://lear.inrialpes.fr/pubs/2010/Kla10.

[3] J. LIEBELT. *Synthetic 3D model-based object class detection and pose estimation*, Université de Grenoble, October 2010, http://lear.inrialpes.fr/pubs/2010/Lie10/.

Articles in International Peer-Reviewed Journal

[4] H. CEVIKALP, D. LARLUS, B. TRIGGS, M. NEAMTU, F. JURIE. *Manifold based local classifiers: linear and non linear approaches*, in "Journal of Signal Processing Systems", October 2010, vol. 61, n^{o} 1, p. 61–73, http://lear.inrialpes.fr/pubs/2010/CLTNJ10.

[5] M. DOUZE, H. JÉGOU, C. SCHMID. *An image-based approach to video copy detection with spatio-temporal post-filtering*, in "IEEE Transactions on Multimedia", March 2010, vol. 12, n^{o} 4, p. 257 - 266, http://hal.inria.fr/inria-00514754.

[6] V. FERRARI, F. JURIE, C. SCHMID. *From images to shape models for object detection*, in "International Journal of Computer Vision", May 2010, vol. 87, n^{o} 3, p. 284–303, http://lear.inrialpes.fr/pubs/2010/FJS10.

[7] M. GUILLAUMIN, T. MENSINK, J. VERBEEK, C. SCHMID. *Face recognition from caption-based supervision*, in "International Journal of Computer Vision", 2010, to appear.

[8] H. JÉGOU, M. DOUZE, C. SCHMID. *Product quantization for nearest neighbor search*, in "IEEE Transactions on Pattern Analysis & Machine Intelligence", 2010, to appear, http://lear.inrialpes.fr/pubs/2011/JDS11/jegou_searching_with_quantization.pdf.

[9] H. JÉGOU, M. DOUZE, C. SCHMID. *Improving bag-of-features for large scale image search*, in "International Journal of Computer Vision", February 2010, vol. 87, n^{o} 3, p. 316-336, http://hal.inria.fr/inria-00514760.

[10] H. JÉGOU, C. SCHMID, H. HARZALLAH, J. VERBEEK. *Accurate image search using the contextual dissimilarity measure*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", January 2010, vol. 32, n^{o} 1, p. 2–11, http://hal.inria.fr/inria-00439311.

[11] D. LARLUS, J. VERBEEK, F. JURIE. *Category level object segmentation by combining bag-of-words models with Dirichlet processes and random fields*, in "International Journal of Computer Vision", June 2010, vol. 88, n^{o} 2, p. 238–253, http://lear.inrialpes.fr/pubs/2010/LVJ10.

[12] D. WEINLAND, R. RONFARD, E. BOYER. *A survey of vision-based methods for action representation, segmentation and recognition*, in "Computer Vision and Image Understanding", 2010, to appear, http://lear.inrialpes.fr/pubs/2010/WRB10a.

Invited Conferences

[13] J. VERBEEK, M. GUILLAUMIN, T. MENSINK, C. SCHMID. *Image annotation with TagProp on the MIR-FLICKR set*, in "ACM Multimedia Information Retrieval", March 2010, p. 537–546, http://lear.inrialpes.fr/pubs/2010/VGMS10.

International Peer-Reviewed Conference/Proceedings

[14] D. ALDAVERT, A. RAMISA, R. TOLEDO, R. L. DE MANTARAS. *Fast and robust object segmentation with the integral linear classifier*, in "IEEE Conference on Computer Vision & Pattern Recognition", June 2010, p. 1046–1053, http://lear.inrialpes.fr/pubs/2010/ARTL10.

[15] M. DOUZE, H. JÉGOU, C. SCHMID, P. PÉREZ. *Compact video description with precise temporal alignment*, in "European Conference on Computer Vision", September 2010, p. 522–535, http://lear.inrialpes.fr/pubs/2010/DJSP10.

[16] M. GUILLAUMIN, J. VERBEEK, C. SCHMID. *Multimodal semi-supervised learning for image classification*, in "IEEE Conference on Computer Vision & Pattern Recognition", June 2010, p. 902 - 909, http://lear.inrialpes.fr/pubs/2010/GVS10.

[17] M. GUILLAUMIN, J. VERBEEK, C. SCHMID. *Multiple instance metric learning from automatically labeled bags of faces*, in "European Conference on Computer Vision", September 2010, p. 634–647, http://lear.inrialpes.fr/pubs/2010/GVS10a.

[18] H. JÉGOU, M. DOUZE, G. GRAVIER, C. SCHMID, P. GROS. *INRIA LEAR-TEXMEX: Video copy detection task*, in "TRECVID Workshop", November 2010.

[19] H. JÉGOU, M. DOUZE, C. SCHMID, P. PÉREZ. *Aggregating local descriptors into a compact image representation*, in "IEEE Conference on Computer Vision & Pattern Recognition", June 2010, p. 3304–3311, http://lear.inrialpes.fr/pubs/2010/JDSP10.

[20] A. KLÄSER, M. MARSZAŁEK, C. SCHMID, A. ZISSERMAN. *Human focused action localization in video*, in "International Workshop on Sign, Gesture, and Activity (SGA) in conjunction with ECCV", September 2010, http://hal.inria.fr/inria-00514845.

[21] J. KRAPAC, M. ALLAN, J. VERBEEK, F. JURIE. *Improving web-image search results using query-relative classifiers*, in "IEEE Conference on Computer Vision & Pattern Recognition", June 2010, p. 1094–1101, http://lear.inrialpes.fr/pubs/2010/KAVJ10.

[22] J. LIEBELT, C. SCHMID. *Multi-view object class detection with a 3D geometric model*, in "IEEE Conference on Computer Vision & Pattern Recognition", June 2010, p. 1688–1695, http://lear.inrialpes.fr/pubs/2010/LS10.

[23] T. MENSINK, G. CSURKA, F. PERRONNIN, J. SÁNCHEZ, J. VERBEEK. *LEAR and XRCE's participation to visual concept detection task - ImageCLEF 2010*, in "Working Notes for the CLEF 2010 Workshop", September 2010, 48, http://lear.inrialpes.fr/pubs/2010/MCPSV10.

[24] T. MENSINK, J. VERBEEK, G. CSURKA. *Trans media relevance feedback for image autoannotation*, in "British Machine Vision Conference", August 2010, p. 20.1–20.12, http://lear.inrialpes.fr/pubs/2010/MVC10.

[25] T. MENSINK, J. VERBEEK, B. KAPPEN. *EP for efficient stochastic control with obstacles*, in "European Conference on Artificial Intelligence", August 2010, p. 675–680, http://lear.inrialpes.fr/pubs/2010/MVK10.

[26] F. PERRONNIN, J. SÁNCHEZ, T. MENSINK. *Improving the Fisher kernel for large-scale image classification*, in "European Conference on Computer Vision", September 2010, p. 143–156, http://lear.inrialpes.fr/pubs/2010/PSM10.

[27] H. SANDHAWALIA, H. JÉGOU. *Searching with expectations*, in "IEEE International Conference on Acoustics Speech and Signal Processing", Signal Processing, IEEE, March 2010, p. 1242–1245, http://lear.inrialpes.fr/pubs/2010/SJ10.

National Peer-Reviewed Conference/Proceedings

[28] M. GUILLAUMIN, J. VERBEEK, C. SCHMID. *Apprentissage de distance pour l'annotation d'images par plus proches voisins*, in "Reconnaissance des Formes et Intelligence Artificielle", France Caen, January 2010, http://hal.inria.fr/inria-00439309.

[29] H. JÉGOU, M. DOUZE, C. SCHMID. *Représentation compacte des sacs de mots pour l'indexation d'images*, in "Reconnaissance des Formes et Intelligence Artificielle", Januanry 2010, http://lear.inrialpes.fr/pubs/2010/JDS10.

Scientific Books (or Scientific Book chapters)

[30] R. RONFARD, G. TAUBIN. *Image and geometry processing for 3D cinematography*, Geometry and Computing, Springer, July 2010, vol. 5, http://lear.inrialpes.fr/pubs/2010/RT10.

Research Reports

[31] M. GUILLAUMIN, T. MENSINK, J. VERBEEK, C. SCHMID. *Face recognition from caption-based supervision*, INRIA, September 2010, RT-392, http://hal.inria.fr/inria-00522185.

[32] A. KLÄSER, M. MARSZAŁEK, I. LAPTEV, C. SCHMID. *Will person detection help bag-of-features action recognition?*, INRIA, September 2010, RR-7373, http://hal.inria.fr/inria-00514828.

[33] A. PREST, C. SCHMID, V. FERRARI. *Weakly supervised learning of interactions between humans and objects*, INRIA, September 2010, RT-391, http://hal.inria.fr/inria-00516477.

[34] D. WEINLAND, R. RONFARD, E. BOYER. *A survey of vision-based methods for action representation, segmentation and recognition*, INRIA, February 2010, RR-7212, http://hal.inria.fr/inria-00459653.