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 Scientist

Cordelia Schmid [Team Leader, INRIA Research Director, DR1, HdR]
Hervé Jégou [INRIA Researcher, CR1, until August '09 at Grenoble, since September '09 at Rennes]
Rémi Ronfard [INRIA Researcher, CR1, since April '09, HdR]
Jakob Verbeek [INRIA Researcher, CR2]

Faculty Member

Roger Mohr [Professor émérite at ENSIMAG, HdR]
Laurent Zwald [Associate professor at UJF, part-time with LJK-SMS]

External Collaborator

Frédéric Jurie [Professor at University of Caen, HdR]

Technical Staff

Matthijs Douze [January '05 – December '10, QUAERO project from August '08]
Benoit Mordelet [GRAVIT project Sep. '07 – Aug. '08, ANR project RAFFUT Sep. '08 – Dec. '09]
Christophe Smekens [INRIA, ODL, September '07 – February '09]

PhD Student

Adrien Gaidon [INPG, Microsoft/INRIA project, October '08 – September '11]
Matthieu Guillaumin [INPG, Ministry of research grant, September '07 – August '10]
Hedi Harzallah [INPG, MBDA project, August '07 – July '10]
Alexander Kläser [INPG, EU project CLASS, November '06 – January '10]
Josip Krapac [University of Caen, ANR project R2I, co-supervision with F. Jurie, Jan. '08 – Dec. '10]
Joerg Liebelt [INPG, EADS scholarship, co-supervision with R. Westermann, TU Munich, since Oct. '06]
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, since Jun. '09]
Gaurav Sharma [University of Caen, ANR project SCARFACE, co-supervision with F. Jurie, since Oct. '09]

Post-Doctoral Fellow

Moray Allan [EU project CLASS December '07 – March '09, ANR project R2I April '09 – December '09]
Tingting Jiang [INRIA December '07 – November '09]
Arnau Ramisa [IIIA-CSIC September '09 – December '09, QUAERO project, starting January '10]
Oksana Yakhnenko [INRIA, ANR project GAIA, starting November '09]

Administrative Assistant

Anne Pasteur [Secretary INRIA]

Other

Nicolas Breitner [Student Intern, INPG ENSIMAG, June '09 – August '09]
Lucie Daubigney [Student Intern, INPG PHELMA, June '09 – August '09]
Gaspard Jakowiak [Student Intern, INPG ENSIMAG, January '09 – May '09]
Javier Montoya-Zegarra [Student Intern, Master-2 Informatics Grenoble, September '09 – August '10]
Harsimrat Sandhawalia [Student Intern, Master-2 Informatics Grenoble, December '08 – January '10]
Heng Wang [Visiting PhD student, LIAMA, Inst. of Automation, Chinese Ac. of Science, Feb. '09 – Jul. '09]

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:

- **Image features and descriptors and robust correspondence.** 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 give better characterizations of visual object classes, for example based on 2D shape descriptors or 3D object category representations, and at defining more powerful measures for visual salience, similarity, correspondence and spatial relations.
- **Statistical modeling and machine learning for visual recognition.** Our work on statistical modeling and machine learning is aimed mainly at making them more applicable to 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 recognition.** Visual recognition requires the construction of exploitable visual models of particular objects and of object and scene 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 diagnostics.
- **Video interpretation.** Humans and their activities are one of the most frequent and interesting subjects of 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 visual shape descriptors to characterize humans and their movements with little or no manual modeling. Video, furthermore, permits to easily acquire large quantities of image data often associated with text. This data needs to be handled efficiently: we need to develop adequate data structures; text classification can help to select relevant parts of the video.

2.2. Highlights of the year

- **Excellent results in ImageCLEF evaluation campaigns**. LEAR participated in the Photo Annotation and Photo Retrieval tasks of the ImageCLEF 2009 evaluation campaign, which is a part of the Cross Language Evaluation Forum (CLEF). In the first task images have to be annotated automatically with relevant concept names, and in the second relevant images have to be retrieved from a set of 500.000 images given a query image and keywords. For both tasks our results obtained a second place among 19 international participating research teams from industry and academia. CLEF is an activity of the TrebleCLEF Coordination Action under the Seventh Framework Programme of the European Commission. See also http://imageclef.org/2009.

- **Action recognition in video**. LEAR has recently developed several successful methods for action recognition in video. A method based on bags of spatio-temporal interest points [24] achieves excellent results in combination with text-based search for retrieving actions [14] as well as for learning the scene context of actions [21]. Furthermore, to localize human actions we have developed a human-centric approach. We first extract human tracks and then characterize the actions with 3D histogram-of-gradient descriptors (Sec. 6.4.2). This allows to precisely localize human actions in space and time. The interaction with objects (Sec. 6.2.5) can further refine the action description.

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 name is by analogy with "bag-of-words" representations in document analysis. The local features are thus sometimes called "visual words"). 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 machines. 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 the EU FP6 project CLASS we investigated methods for visual learning with little or no manual labeling and semantic-level image and video querying. 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.

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

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. The ANR RAFFUT project investigates the problem of content protection for videos. Another type of application is the detection of specific content from images and videos, which can be used for a large number of problems. Transfer to such problems is the goal of the start-up MilPix, to which our current technologies for image search are licenced.

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 indexing

Participants: Matthijs Douze, Hervé Jégou, Benoit Mordelet, Cordelia Schmid.

Our large-scale image indexing software has been improved in 2009 as follows:

1. the point-based scoring method now includes the burstiness measure [19]
2. several algorithms that index global descriptors have been added. The first one builds an inverted file from GIST descriptors [13], the second one indexes several "MiniBofs" [20].
3. An efficient high-dimensional nearest-neighbor method has been added: the product quantizer [29].
4. A client-server version of the video search engine is now included. It is aimed at application scenarios where the server should not have access to the unprocessed video data.
5. The on-line demonstrator now indexes 10 million images (2 million in the 2008). It can tested at http://bigimbaz.inrialpes.fr.

Version 2.4 of LEAR's image search engine, BIGIMBAZ, has been registered at the "Agence pour la Protection des Programmes", under IDDN.FR.001.510004.001.S.A.2008.000.21000. It has been transferred for research purposes only to Stanford University, San Diego University, and the California Institute of Technology.

We have also adapted and tested the software for evaluation by the i-Dash European project (http://www.i-dash.eu/). This project aims at producing a platform that the police can use in their investigations which involve large quantities of child abuse video material. Our video indexing system can be used to recognize strongly distorted and low-quality video clips. It has been evaluated and shortlisted by TNO (http://www.tno.nl/) to detect known instances of illegal videos.

5.2. Datasets

Participants: Moray Allan, Matthijs Douze, Matthieu Guillaumin, Hervé Jégou, Cordelia Schmid, Jakob Verbeek.

Relevant datasets are important to assess recognition methods. They allow to point out the weakness of existing methods and push forward the state-of-the-art. Datasets should capture a large variety of situations and conditions. Benchmarking procedures allow to compare the strengths of different approaches and provide clear and broadly understood performance measures.

In addition to the datasets we previously created, we released several new datasets this year as well as pre-processed image descriptors. Our publicly accessible datasets are available at http://lear.inrialpes.fr/data.

Annotated Flickr image data set

This new data set contains Flickr images and user annotations for 20 object categories and 20 combinations of object categories. For each object category images were downloaded from the Flickr website, and ranked using our method described in [11]. For evaluation images that were top-ranked by our method have been manually annotated to indicate whether they contain the object category. In total the data set contains about 100.000 images.

Image features for image annotation data sets

A fair comparison of image annotation methods, or machine learning methods in general, requires that the same feature set is used. In this way we can separate the contributions due to good image features from those due to good learning methods. In recent work [15] we presented state-of-the-art performance on three benchmark datasets for image annotation. We released the image features computed for these datasets (45.000 images in total) to allow direct comparison with our results.

Hollywood-2 Human Actions and Scenes Dataset

The Hollywood-2 Human Actions and Scenes Dataset [21] contains 12 classes of human actions and 10 classes of scenes distributed over 3669 video clips and approximately 20.1 hours of video. The dataset provides a benchmark for human action recognition in realistic and challenging settings; it is composed of video clips from 69 movies divided into 33 training and 36 test movies. There are two training sets: an automatic and a clean one. The automatic one is obtained using automatic script-based action annotation and contains 810 video samples with approximately 60% correct labels. The clean set contains 823 video samples with manual labels. The action test set contains 884 manually annotated samples. Scene classes are selected automatically from scripts such as to maximize co-occurrence with the given action classes and to capture action context as described in [21]. Scene video samples are then generated using script-to-video alignment. The labels of test scene samples are manually verified to be correct.

Copydays

Copydays is an image dataset designed to evaluate copy detection systems. It comprises a reference set of 157 personal holidays photos. These images were transformed using three kinds of artificial attacks: JPEG compression, cropping (extracting subparts) and "strong" (print and scan, paint, change in contrast, perspective effect, blur, very strong crop, etc.). This dataset was merged in [13] with a very large image set (up to 110 million images) to evaluate the behavior of different indexing schemes for copy detection on a large scale. Sample images and transformations are illustrated in Fig 2.

6. New Results

6.1. Large-scale image search

6.1.1. Burstiness of visual elements

Participants: Matthijs Douze, Hervé Jégou, Cordelia Schmid.

Burstiness, a phenomenon initially observed in text retrieval, is the property that a given element appears more times in a document than a statistically independent model would predict. We have shown that burstiness translates to visual words in images [19], see Figure 1 for an illustration. One can observe that many detected regions are assigned to the same visual word. The examples include man-made objects such as buildings, church windows and playing cards as well as textures such as a brick wall and corals. In both cases the repetitiveness stems from the scene property, for example the windows of the buildings are very similar and the bricks are repeated.

Figure 1. Illustration of burstiness. Features assigned to the most "bursty" visual word of each image are displayed.

In the context of image search, burstiness corrupts the visual similarity measure, i.e., the scores used to rank the images. We, therefore, proposed a strategy to handle visual bursts for bag-of-features based image search systems. Experimental results on three reference datasets show that handling burstiness with the proposed method significantly and consistently outperforms the state of the art.

6.1.2. Compact representation of bag-of-features

Participants: Matthijs Douze, Hervé Jégou, Cordelia Schmid.

One of the main limitations of image search based on bag-of-features is the memory usage per image. Only a few million images can be accessed on a single machine in quasi real-time. In [20], [26] we first evaluated how the memory usage is reduced by using lossless index compression. We then proposed an approximate representation of bag-of-features obtained by projecting the corresponding histogram onto a set of pre-defined sparse projection functions, producing several image descriptors. Coupled with a appropriate indexing structure, an image is represented by a few hundred bytes. A distance expectation criterion is then used for ranking images. Our method is at least one order of magnitude faster than standard bag-of-features while providing excellent search quality.

6.1.3. Approximate nearest neighbor search with quantization

Participants: Matthijs Douze, Hervé Jégou, Harsimrat Sandhawalia, Cordelia Schmid.

We have proposed two approaches for nearest neighbor search in the presence of severe memory constraints. The key idea is to see the problem of search as a distance estimation problem. Our first approach [22] mimics a source coding approach, and formulates the problem of generating compact signature as a rate-distortion problem. In the spirit of source coding algorithms, we aim at minimizing the reconstruction error on the squared distances with a constraint on the memory usage. The vectors are ranked based on their distance expectations to the query vector.

The idea is pushed further in [29], where vector quantization based on a product quantizer is used to obtain a distance estimation. The method is advantageously used in an asymmetric manner, by computing the distance between a vector and code. This is in contrast to competing techniques such as spectral hashing that only compare codes. The method is shown to outperform two state-of-the-art approaches of the literature. Timings measured when searching a vector set of 2 billion vectors are shown to be excellent given the high accuracy of the method.

6.1.4. Evaluation of GIST descriptors for web-scale image search

Participants: Laurent Amsaleg [CNRS - IRISA], Matthijs Douze, Hervé Jégou, Harsimrat Sandhawalia, Cordelia Schmid.

We have evaluated the search accuracy and complexity of the global GIST descriptor [13] for two applications, where a local description is generally preferred: same location/object recognition and copy detection. We also proposed an indexing strategy for global descriptors that optimizes the trade-off between memory usage and precision. Our scheme provides a reasonable accuracy in some widespread application cases together with very high efficiency: In our experiments, querying an image database of 110 million images takes 0.18 second per image on a single machine. For common copyright attacks, this efficiency is obtained without noticeably sacrificing the search accuracy compared with state-of-the-art approaches. See Figure 2 for example queries and search results.

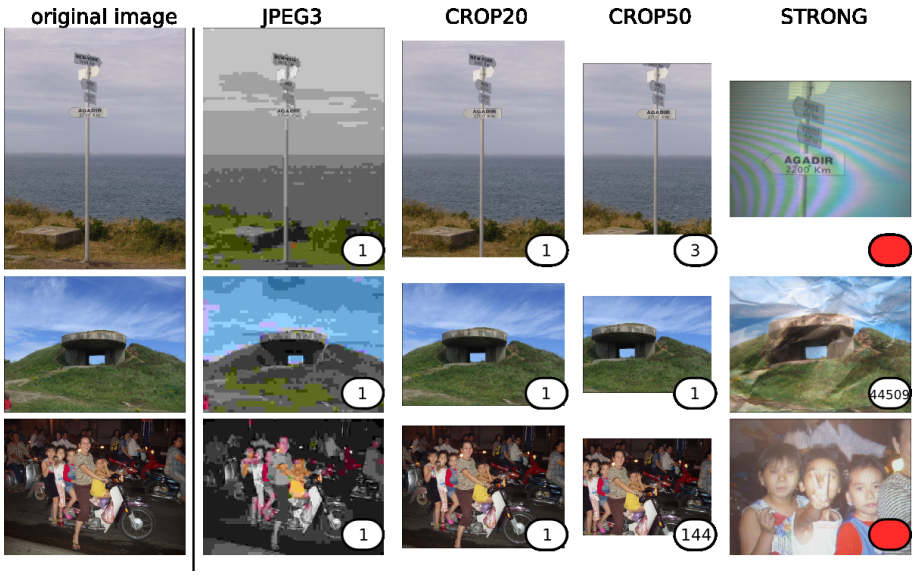


Figure 2. We search images in a 110 million-image dataset using the GIST descriptor and our large-scale indexing approach. Query images (right) are degraded more or less severely. The numbers indicate the rank of the original image (left) in the resulting response list. Results are excellent for JPEG3 and CROP20 and very good for CROP50. For "strong" transformations two out of three examples were not found (plain circle).

6.1.5. Aggregating local descriptors into a compact image representation

Participants: Matthijs Douze, Hervé Jégou, Patrick Pérez [INRIA Rennes], 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. To address this problem, we first propose a simplification of the Fischer Kernel image representation, which is a way of aggregating local image descriptors into a vector of limited dimension. We then present an approach for coding and indexing such vectors that preserves well the accuracy of the vectorial Euclidean 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 requiring 20 bytes of memory.

6.2. Semi-supervised learning and structuring of visual models

6.2.1. Automatic image annotation

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

Image auto-annotation is an important open problem in computer vision. For this task we developed TagProp, a discriminatively trained nearest neighbor model [15], [25]. In TagProp, tags of test images are predicted using a weighted nearest-neighbor model to exploit labeled training images. Neighbor weights are based on neighbor rank or distance. TagProp allows the integration of metric learning by directly maximizing the log-likelihood of the tag predictions in the training set. In this manner, we can optimally combine a collection of image similarity metrics that cover different aspects of image content, such as local shape descriptors, or global color histograms. We also introduced a word specific sigmoidal modulation of the weighted neighbor tag predictions to boost recall of rare words.

We investigated the performance of different variants of our model and compared to existing work. We presented experimental results for three challenging data sets. On all three, TagProp has proven to significantly improvement over the current state-of-the-art. In a follow-up paper [23] we make an extensive comparison to results obtained with support vector machine classifiers, and consider image categorization problems where images are accompanied by user-generated keywords. An on-line demonstration of our system is available at http://pascal.inrialpes.fr/local/tagprop/. Figure 3 shows several example images with the automatically generated annotations.

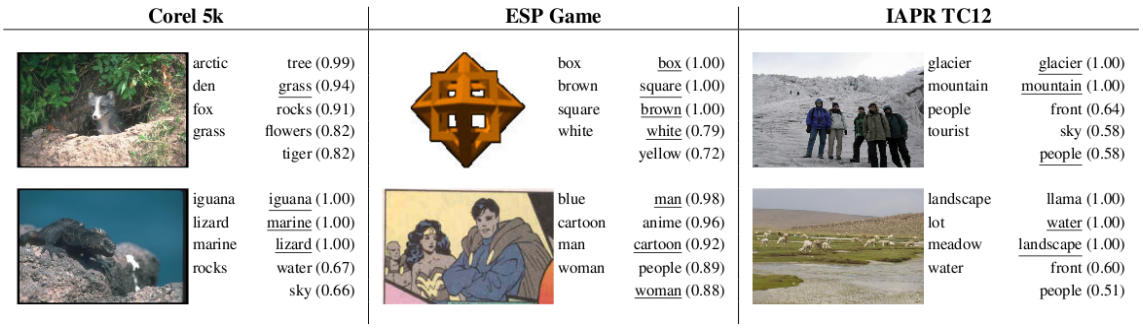


Figure 3. Two example images from the three data sets that we used in our experiments. For each image we show the manual annotation (left) and the automatically predicted one (right), where we give the confidence value for each predicted word, and underline it when it appears in the manual annotation.

6.2.2. Semi-supervised image categorization

Participants: Matthieu Guillaumin, Cordelia Schmid, Jakob Verbeek.

In on-going work, we study the problem of image categorization using semi-supervised techniques. The goal is to decide if an image belongs to a certain category or not. In the standard supervised setting, a binary classifier is learned from manually labeled images. Using more labeled examples typically improves performance, but obtaining the image labels is a time consuming process. We are interested in how other sources of information can aid the learning process given a fixed amount of labeled images. In particular, we consider a scenario where keywords are associated with the training images, e.g. as found on photo sharing websites. The goal is to learn a classifier for images alone, but we will use the keywords associated with labeled and unlabeled images to improve the classifier using semi-supervised learning. We learn a first strong classifier using both the image content and keywords, and use it to predict the labels of unlabeled images. We then learn a second classifier that only takes visual features as input, and train it from the labeled images and the output of the first

classifier for the unlabeled ones. 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. We consider a scenario where we do not use any manual labeling, but directly learn classifiers from the image tags. Also in this case using the semi-supervised approach improves classification performance. Figure 4 shows example images with their keywords and category labels as used in our experiments.

Figure 4. Example images used in our experiments, together with the user tags and category labels.

6.2.3. Ranking user-annotated images for multiple query terms

Participants: Moray Allan, Jakob Verbeek.

In [11] we considered how image search on photo-sharing sites like Flickr can be improved by taking into account the users who provided different images. Further we showed that, when searching for multiple terms, performance can be increased by learning a new combined model and taking account of images which partially match the query. Search queries are answered by using a mixture of kernel density estimators to rank the visual content of images from the Flickr website whose noisy tag annotations match the given query terms. Experiments show that requiring agreement between images from different users allows a better model of the visual class to be learnt, and that precision can be increased by rejecting images from 'untrustworthy' users. We focus on search queries for multiple terms, and demonstrate enhanced performance by learning a single model for the overall query, treating images which only satisfy a subset of the search terms as negative training examples. Figure 5 shows some Flickr images returned by a textual search for 'boat', and the highest-ranking results for these queries according to our model.

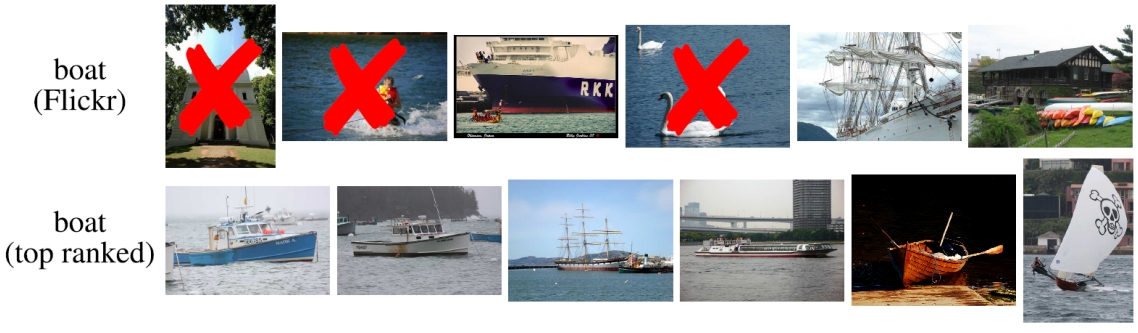


Figure 5. Images matching the query 'boat' (cross marks irrelevant ones), and those top ranked by our model.

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

Participants: Moray Allan, Frédéric Jurie, Josip Krapac, Jakob Verbeek.

In this on-going work we propose an image re-ranking method, based on textual and visual features, that does not require learning a separate model for every new search query. 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 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.

The second contribution of this work is that we present a new public data set of images returned by a web search engine for 353 search queries, along with their associated meta-data, and ground-truth annotations for all images. We hope that this new data set will facilitate further research in improving image search.

As an example, Figure 6 shows the top-ranked images given by a web search engine for the query 'Eiffel tower' (top), and the top-ranked images after re-ranking by our proposed classification method (bottom).

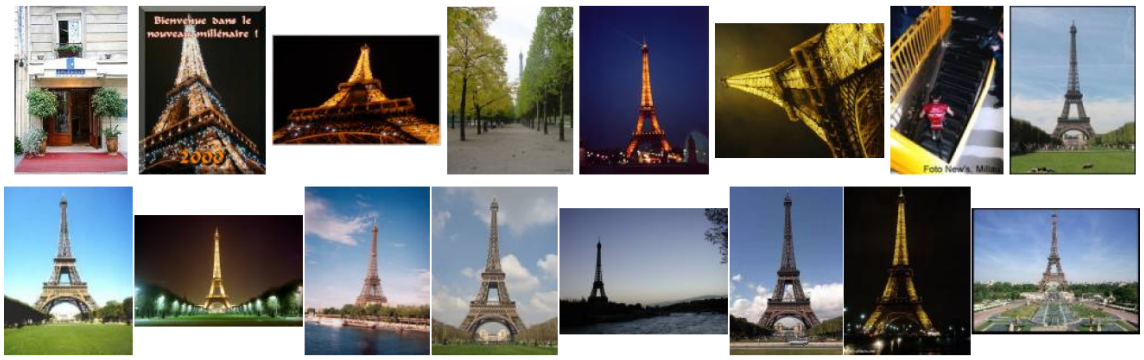


Figure 6. Example images for the query 'Eiffel tower', the top-ranked images given by a web search engine for the query (top), and the top-ranked images after re-ranking by our proposed classification method (bottom).

6.2.5. Automatic learning of interactions between humans and objects

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

In this on-going work we introduce a novel human-centric approach for learning human actions modeled as interactions with objects. Interactions are often the main characteristic of an action, see Figure 7. The action 'playing trumpet' can be described as a human holding a trumpet in a certain position. Characteristic features are the object trumpet and its spatial relation to the human.

Our approach first detects humans with a part-based human detector able to cope with various degrees of visibility: we map different part detections to a common reference frame and assign them a single score. We then use a learning algorithm to determine the action object and the spatial relation of that object to the human. Starting from a set of images depicting an action, our method produces a probabilistic model of the human-object interaction, i.e., it determines automatically the relevant object and the spatial relation between the human and the object. Images for the actions 'playing a trumpet', 'riding a bike', and 'wearing a hat' were obtained via Google Images search and text search on the IAPR TC-12 dataset. Results show that humans, action objects, as well as spatial human-object interactions can be determined automatically, see Figure 7. Note that our approach can handle images not containing the action, as may occur in the case of keyword based text search.

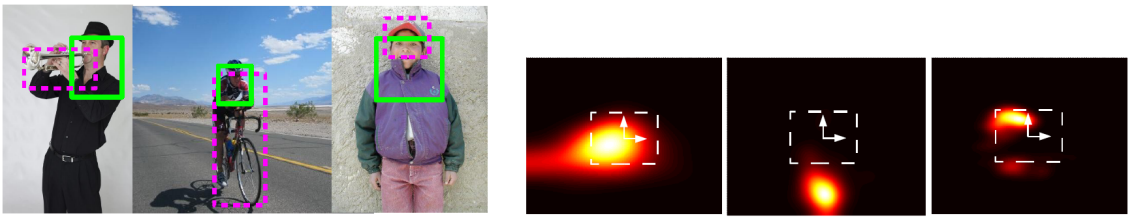


Figure 7. Left: Example results showing the automatically detected human (green) and object (pink) for the actions 'playing trumpet', 'riding bike' and 'wearing hat'. Right: Human-object spatial relations learned by our method. The reference frame is based on the human detection. The trumpet (left plot) is in front of the human, the bike (center plot) below the person, and the hat (right plot) on top of the person.

6.3. Supervised methods for visual object recognition and localization

6.3.1. Learning metrics for visual identification

Participants: Matthieu Guillaumin, Cordelia Schmid, Jakob Verbeek.

Face identification, determining whether two face images depict the same person or not, is difficult due to variations in scale, pose, lighting, background, expression, hairstyle, and glasses. We have introduced two methods for learning robust distance measures [16]: (a) a logistic discriminant approach which learns a Mahanalobis metric from a set of labelled image pairs, and (b) a nearest neighbour approach which computes the probability for two face images to belong to the same person. We evaluated our approaches on the Labeled Faces in the Wild data set, a large and very challenging data set of faces from Yahoo!News. The evaluation protocol for this data set defines a restricted setting, where a fixed set of positive and negative image pairs is given, as well as an unrestricted one, where faces are labelled by their identity. At time of submission, we were the first to present results for the unrestricted setting, and showed that our methods benefit from this richer training data, much more so than the current state-of-the-art method. Our results of 79.3% and 87.5% correct for the restricted and unrestricted setting respectively, significantly improved over the current state-of-the-art result of 78.5%. Confidence scores obtained for face identification were also used for applications like clustering or recognition from a single training example. We showed that our learned metrics also improve performance for these tasks. See Figure 8 for two example face clusters.

Figure 8. Examples face-clusters found using our method. The left cluster contains faces of only one person despite the variations in pose and expression within each cluster, the right cluster contains two faces of another person.

6.3.2. Combining efficient object localization and image classification

Participants: Hedi Harzallah, Frédéric Jurie, Cordelia Schmid.

We have developed a unified approach for object localization and classification [17]. Objects are localized with an efficient two stage sliding window method that combines the efficiency of a linear classifier with the robustness of a sophisticated non-linear one. The first stage generates object hypotheses based on a sliding window and a linear support vector machine (SVM) classifier. This allows to rapidly reject most negative windows. The remaining one are then re-scored with a non-linear SVM classifier with a χ^2 radial basis function (RBF) kernel. This allows to significantly improve the localization results, see Figure 9 for example detections.

Given object detections, we show that a contextual combination with image classification can improve detection results further. Even though the two tasks are different, they are obviously related. Knowing the class of an image can help to detect hardly visible objects. Indeed, if an object is only partially visible it will be hard to find by the detector while the classifier could still have enough information (context, object parts) to decide on the presence of the object.Experimental results show that our combined object localization and classification methods outperform the state-of-the-art on the PASCAL VOC 2007 and 2008 datasets. Our approach also gives very good results on a dataset (infrared images) provided by MBDA.

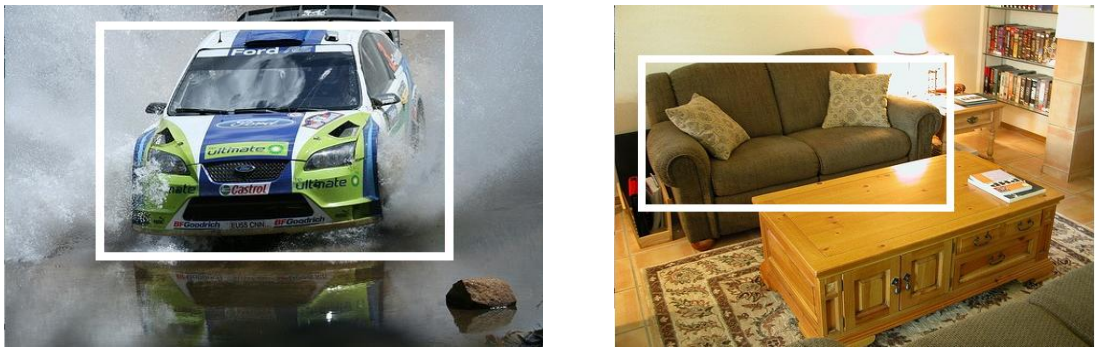


Figure 9. Example detections produced by our combined classification and detection method of cars and sofas on images of the PASCAL VOC dataset.

6.3.3. Learning shape models for object matching

Participants: Tingting Jiang, Frédéric Jurie, Cordelia Schmid.

The aim of this work [18] is to learn an a-priori shape model for an object class and to improve shape matching with the learned shape prior. Given images of example instances, we learn a mean shape of the object class as well as variations of non-affine and affine transformations separately based on the thin plate spline (TPS) parametrization.

Unlike previous methods, for learning, we represent shapes by vector fields instead of features which makes our learning approach general. During shape matching, we inject the shape prior knowledge and make the matching result consistent with the training examples. This is achieved by an extension of the TPS-RPM algorithm which finds a closed form solution for the TPS transformation coherent with the learned transformations.

We tested our approach by using it to learn shape prior models for all the five object classes in the ETHZ Shape Classes data set. The results show that more accurate shape models are learned than in previous work, and the learned shape models improve object classification.

6.3.4. Region-based image segmentation

Participants: Tingting Jiang, Frédéric Jurie, Cordelia Schmid.

The aim of this work is to combine state-of-art image segmentation and object detection methods. We first developed a region classifier based on color and texture features. For color features, we applied k-means on the RGB values of pixels to generate a code book of 200 words. For texture, we used SIFT and there are 4000 words in the code book.

During training, for each object region (manually annotated), we concatenated the histograms of color and texture. For each class, we trained a nonlinear SVM classifier with an intersection kernel. During testing, we first generated an over-segmentation using the Berkeley Segmentation Engine and then used the learned classifiers to estimate the class membership probabilities for each region. After the preliminary segmentation process, we combined the segmentation results with that of the object detector.

Our results were among the top ones in the PASCAL VOC 2009 image segmentation challenge.

6.3.5. 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. Appearance and geometry are treated as separate learning tasks with different training data.

Our approach uses a part model which discriminatively learns the object appearance with spatial pyramids from a database of real images, and encodes the 3D geometry of the object class with a generative representation built from a database of synthetic models. The geometric information is related to the 2D training data based on viewpoint annotations and allows to perform an approximate 3D pose estimation for generic object classes. The pose estimation provides an efficient method to evaluate the likelihood of groups of 2D part detections with respect to a full 3D geometry model in order to disambiguate and prune 2D detections and to handle occlusions.

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 the Stanford 3D car and bicycle databases, the current state-of-the-art benchmark for 3D object detection, our approach outperforms previously published results for viewpoint estimation. See Figure 10 for an illustration.

Figure 10. From left to right: a pre-detection, detected parts, initial orientation estimate, and final pose estimate.

6.4. Action recognition in video

6.4.1. Evaluation of local spatio-temporal features for action recognition

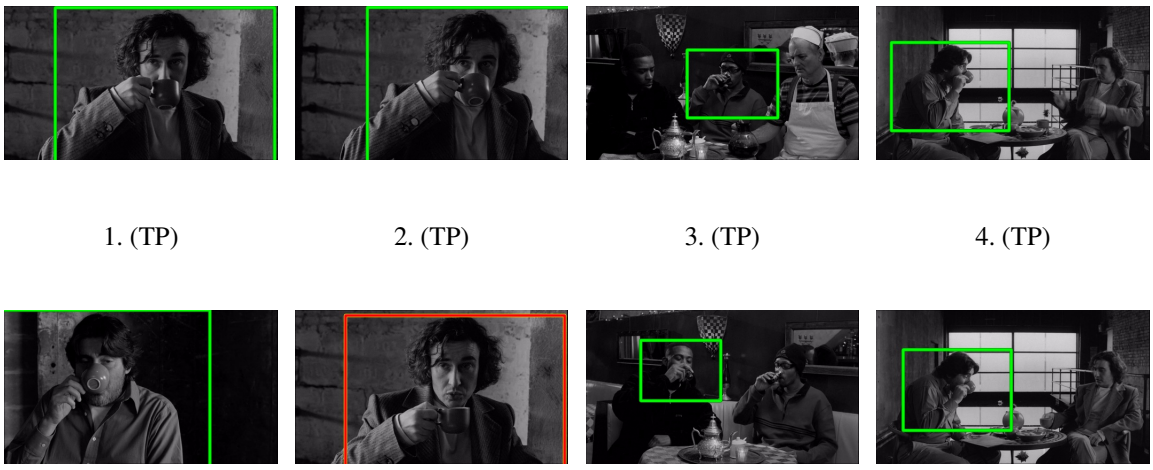
Participants: Alexander Kläser, Ivan Laptev [INRIA Rocquencourt], Cordelia Schmid, Muhammad Ullah [INRIA Rennes], Heng Wang.

Local space-time features have recently become a popular video representation for action recognition. Several methods for feature localization and description have been proposed in the literature, and promising recognition results were demonstrated for different action datasets. The comparison of those methods, however, is limited given the different experimental settings and various recognition methods used. In our current work [24], we carried out an extensive evaluation of local spatio-temporal features. We defined a common evaluation framework based on bag-of-features video sequence classification. Experiments showed that dense sampling consistently outperforms all tested methods for feature localization in realistic video settings. Note, however, that dense sampling also produces a very large number of features. Among the different feature point detectors, we observe a similar performance. For the tested feature descriptors, the combination of gradient based and optical flow based descriptors seems to be a good choice for action recognition. The combination of dense sampling with the HOG/HOF descriptor provides best results for the most challenging dataset Hollywood2. On the UCF dataset, the HOG3D descriptor performs best in combination with dense sampling.

6.4.2. Human focused action localization in video

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

Early work on action recognition in video used sequences with static cameras, simple backgrounds and fully visible bodies. Approaches developed in this context were robust to variations in the actor and action, but not to changes of viewpoint, scale or lighting; partial occlusion and varying background. Recent work uses video material from movies, i.e., less controlled and much more challenging data.



1. (TP) 2. (TP) 3. (TP) 4. (TP)

5. (TP) 6. (FP) 7. (TP) 8. (TP)

Figure 11. Top 8 drinking detections for the movie Coffee and Cigarettes, including true positives (TP) and false positives (FP). The detection at rank 6 is incorrect due to its imprecise localization in time.

To localize human actions in such movies, we develop a human-centric approach. Our goal is to localize the action in time through the sequence and spatially in each frame. We first extract spatio-temporal human tracks and then detect actions within these using a sliding window classifier. Our human tracker is able to cope with a wide range of postures, articulations, motions and camera viewpoints. The tracker includes detection interpolation and a principled classification stage to suppress false positives. To localize actions within the extracted tracks, we introduce a spatio-temporal 3D histogram-of-gradient based descriptor adapted to the track. We show that tracks reduce search complexity and can be reused for multiple human actions, without performance loss.

Experimental results are presented for the actions of drinking and smoking on the *Coffee and Cigarettes* dataset, and for phoning and standing-up on the Hollywood2 dataset. We compare with previous methods on this material and demonstrate a significant improvement over the state of the art. Figure 11 shows the top eight drinking detections in approximately 24 minutes of the *Coffee and Cigarettes* movie.

6.4.3. Mining visual actions from movies

Participants: Adrien Gaidon, Marcin Marszałek [University of Oxford], Cordelia Schmid.

In this work [14] we present an approach for mining visual actions from real-world videos. Given a large number of movies, we want to automatically extract short video sequences corresponding to visual human actions. We can then visually discover which actions are performed and also collect training data for action recognition. We first retrieve action sequences corresponding to specific verbs extracted from the transcripts aligned with the videos. Not all of the samples visually characterize the action and, therefore, we rank these videos by visual consistency. Negative samples are obtained by randomly sampling the rest of the videos. We propose a novel ranking algorithm using an iterative re-training scheme for Support Vector Regression machines (SVR) referred to as 'iter-SVR'. Experimental results explore actions in 144 episodes (more than 100 hours) of the TV series "Buffy the Vampire Slayer" and show that our iter-SVR approach outperforms other commonly used approaches. Examples of retrieved actions are shown in Figure 12.

Figure 12. Key frames of the top 5 'walk' and 'kiss' samples and of the first false positive (FP), (for 'walk' at rank 30, and for 'kiss' at rank 37) obtained with our iter-SVR method.

6.4.4. Learning human actions and their context

Participants: Ivan Laptev [INRIA Rocquencourt], Marcin Marszałek [University of Oxford], Cordelia Schmid.

We exploit the context of natural dynamic scenes for human action recognition in video [21]. Human actions are frequently constrained by the purpose and the physical properties of scenes and demonstrate high correlation with particular scene classes. For example, eating often happens in a kitchen while running is more common outdoors, cf. fig. 13. Our contribution is three-fold: (a) we automatically discover relevant

scene classes and their correlation with human actions, (b) we show how to learn selected scene classes from video without manual supervision and (c) we develop a joint framework for action and scene recognition and demonstrate improved recognition of both in natural video.

Our approach uses movie scripts as a means of automatic supervision for training. For selected action classes we identify correlated scene classes in these scripts and then retrieve video samples of actions and scenes for training using script-to-video alignment. Our visual models for scenes and actions are formulated within the bag-of-features framework and are combined in a joint scene-action classifier. We validate the method on a new large dataset with twelve action classes and ten scene classes acquired from 69 movies.

Experimental results demonstrate the gain in performance for action classification when using contextual scene information.

(a) eating, kitchen (b) eating, cafe (c) eating, restaurant

Figure 13. Video samples from our dataset with automatically assigned labels for the action (eating in these examples), and the context.

7. Contracts and Grants with Industry

7.1. Start-up Milpix

Participants: Hervé Jégou, 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 have been licensed to the start-up: BIGIMBAZ and OBSIDIAN. Hervé Jégou and Cordelia Schmid are the scientific advisers of the start-up MILPIX.

7.2. MDBD Aerospatiale

Participants: Hedi Harzallah, Frédéric Jurie, 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) as well as a one year transfer contract on matching and tracking (11/2005-11/2006).

In December 2006 started a contract for three and a half years on object detection in the presence of severe changes of the imaging conditions and if the images of the objects are very small. Our solution is based on designing appropriate image descriptors and using context information to improve the localization performance. The PhD scholarship of Hedi Harzallah which started in August 2007 is funded by this contract.

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

Participants: Adrien Gaidon, Cordelia Schmid.

This collaborative project, starting September 2008, brings together the WILLOW, LEAR, and VISTA project-teams with MSR researchers in 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 we propose will focus on fundamental computer science research in computer vision and machine learning, and its application to archeology, cultural heritage preservation, environmental science, and sociology, and it will be validated by collaborations with researchers and practitioners in these fields. 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.

8. Other Grants and Activities

8.1. National Projects

8.1.1. QUAERO

Participants: Matthijs Douze, Hervé Jégou, Frédéric Jurie, Arnau Ramisa, Harsimrat Sandhawalia, Cordelia Schmid, Jakob Verbeek.

Quaero is a French-German search engine project, supported by OSEO. It runs from 2008 to 2013 and includes many academic and industrial partners, including 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, clustering and search. See http://www.quaero.org for details.

8.1.2. ANR Project GAIA

Participants: Hervé Jégou, Cordelia Schmid, 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.

8.1.3. ANR Project RAFFUT

Participants: Matthijs Douze, Hervé Jégou, Benoit Mordelet, Cordelia Schmid.

RAFFUT is an ANR (Agence Nationale de la Recherche) "audiovisuel et multimédia" project that started in December 2007 for two years. This project aims at detecting pirated videos. The main issues addressed by this project are 1) how to handle the scalability issues that arise when dealing with extremely large datasets ; 2) how to improve the accuracy of the search if the videos have suffered very strong attacks, as for example low-quality camcorded copies of movies. The partners are the company Advestigo (http://www.advestigo.com) and LEAR. Advestigo is one of the leaders in the growing "digital asset management market". Its technology is oriented towards video piracy, in particular for detecting fraudulent content on user-generated websites such as YouTube or DailyMotion.

8.1.4. ANR Project R2I

Participants: Moray Allan, Frédéric Jurie, 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 final goal of this project is a system for interactive search, which can index about one billion of images and provide users with advanced interaction capabilities. 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, both specialists in object recognition.

8.1.5. ANR Project SCARFACE

Participants: Frédéric Jurie, Cordelia Schmid, Gaurav Sharma.

Video surveillance systems are currently installed in many public areas. As their number increases, the manual analysis becomes impossible. The ANR project SCARFACE will develop 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. SCARFACE is a three-year project (2009-2011) funded by the French Research Agency (ANR). The participant are: the university of Lille (investigating legal issues of such tools), the INRIA-Imedia research group (developing image clustering and interactive retrieval algorithms), SpikeNet (developing technology for face/person detection), EADS (developing the experimental framework) as well as the University of Caen and the INRIA-LEAR group (developing vision algorithms for semantically characterizing persons in images and videos).

8.2. International Projects

8.2.1. FP6 European Project CLASS

Participants: Moray Allan, Matthieu Guillaumin, Alexander Kläser, Thomas Mensink, Cordelia Schmid, Jakob Verbeek.

CLASS (Cognitive-Level Annotation using latent Statistical Structure) is a 6th framework Cognitive Systems STREP that started in January 2006 for three and half years. It is a basic research project focused on developing a specific cognitive ability for use in intelligent content analysis: the automatic discovery of content categories and attributes from unstructured content streams. It studies both fully autonomous and semi-supervised methods. The work combines robust computer vision based image descriptors, machine learning based latent structure models, and advanced textual summarization techniques. The potential applications of the basic research results are illustrated by three demonstrators: an image interrogator that interactively answers simple user-defined queries about image content; an automatic annotator for people and actions in situation comedy videos; an automatic news story summarizer. The Class consortium is interdisciplinary, combining leading European research teams in visual recognition, text understanding and summarization, and machine learning: LEAR; LJK; Oxford University, UK; K.U. Leuven, Belgium; University of Helsinki, Finland; and MPI Tuebingen, Germany.

8.2.2. FP7 European Network of Excellence PASCAL 2

Participants: Adrien Gaidon, Matthieu Guillaumin, Frédéric Jurie, 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. Bilateral relationships

8.3.1. Associated team Tethys

Participants: David Forsyth [UIUC], Martial Hebert [CMU], Jean Ponce [ENS Ulm], Cordelia Schmid.

The associated team Tethys started in January 2007 for three years, see http://lear.inrialpes.fr/people/schmid/bilantriennal-tethys.htm. It associates two INRIA project-teams, LEAR and WILLOW, with two teams in the US, at Carnegie Mellon University and at University of Illinois Urbana-Champaign. The topic of this collaboration is visual recognition of objects with an emphasis on 3D representations for recognition and human activity classification in videos. In 2009, several visits of senior and junior researchers took place.

9. Dissemination

9.1. Leadership within the scientific community

- Conference and workshop organization:
  - C. Schmid: Co-organizer of CVPR'09 Workshop on Feature Detectors and Descriptors.
  - C. Schmid: Co-organizer of International Workshop on Video, Barcelona, Spain.
  - F. Jurie: Co-organizer of RFIA'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 chairs:
  - C. Schmid: CVPR 2010.
  - C. Schmid: ECCV 2010.
  - C. Schmid: ICCV 2009.
  - C. Schmid: RFIA 2010.
- Program committees:
  - CVPR'2009: H. Jégou, F. Jurie, C. Schmid, J. Verbeek.
  - ICCV'2009: H. Jégou, F. Jurie, J. Verbeek.
  - NIPS'2009: F. Jurie, C. Schmid, J. Verbeek.
  - RFIA'2010: J. Verbeek.
- Prizes:
  - Our submissions to the ImageCLEF evaluation campaign for the "Photo Annotation" and "Photo Retrieval" tracks obtained a second place among 19 participants for each track, see http://imageclef.org/2009 and our working notes paper [12].
  - M. Guillaumin received a Google Student Grant to attend ICCV'09.
  - J. Verbeek received an Outstanding Reviewer Award at CVPR'2009.
- Other:
  - H. Jégou was a reviewer for the ANR program CONTINT.

- F. Jurie is vice-head of AFRIF (the French section of the IAPR).
- F. Jurie is scientific co-director of GDR ISIS (national interest group on image analysis).
- F. Jurie was a member of the evaluation committee for CSOSG projects of the Agence Nationale de la Recherche, 2009.
- F. Jurie was a member of the recruiting committees at universities of Caen, Rennes and Rouen, 2009.
- R. Mohr is president of Grilog, Grenoble Isère Logiciel, 2007–2009.
- R. Mohr is an invited professor at Australian National University and NICTA, starting October 2009 for 4 months.
- C. Schmid is a member of INRIA's "Commission d'Évaluation". She participated in several recruitment committees in 2009 and was in charge of the CR2/CR1 2009 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)" starting 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. Guillaumin, several exercise classes in the context of a "monitorat" at ENSIMAG, in total 64h.
- M. Douze and H. Jégou, Multimedia Databases, 3rd year ENSIMAG, 16h.
- A. Gaidon, Introduction to Computer Science for High School Students, 6h.
- C. Schmid and J. Verbeek, Machine Learning & Category Representation, Master-2 MoSIG, Univ. Grenoble, 18h.
- C. Schmid, Object recognition and computer vision, Master-2 MVA, ENS ULM, 10h.

9.3. Invited presentations

- M. Douze, *Recherche d'images à grande échelle*, Presentation and demonstration at "Rencontres INRIA-Industrie", Lille, June 2009.
- M. Douze, *Web-scale image search*, journée GDR ISIS "Passage à l'échelle dans la recherche d'information multimédia", ENST Paris, June 2009.
- M. Douze, *Indexation d'images et de vidéos*, Demonstration at Grenoble Innovation Fair, Grenoble, October 2009.
- M. Guillaumin, *Discriminative metric learning in nearest neighbor models for image auto-annotation*, seminar at Xerox Research Centre Europe, Meylan, September 2009.
- H. Harzallah, *Vehicle Identification*, presentation at MCM-ITP conference, Lille, June 2009.
- H. Jégou, *Selected topics in large scale image search*, seminar INRIA Vistas/Fluminance, St Jacut de la mere, May 2009.
- H. Jégou, *Selected topics in large scale image search*, seminar INRIA Texmex, Quiberon, June 2009.
- F. Jurie, *Recent advances in image representation for image segmentation*, IEEE First Workshop on Emergent Issues in Large Amounts of Visual Data (WS-LAVD) in conjunction with ICCV, October 2009.
- F. Jurie, *Recent advances in image representation for image segmentation, object class detection and image classification*, MIA'09 - Mathematics and Image Analysis, December 2009.

- R. Mohr, *Analyse de quelques avancés en vision par ordinateur*, Ecole des Jeunes Chercheurs Orasis'09, June 2009.
- C. Schmid, *Discriminative metric learning in nearest neighbor models for image auto-annotation*, Seminar at CMU, Pittsburgh, September 2009.
- C. Schmid, *Large scale image search*, International Workshop on Recent Trends in Computer Vision, Kyoto, Japan, June 2009.
- C. Schmid, *Learning classes and context of human actions from movies*, International Workshop on Video, Barcelona, Spain, May 2009.
- C. Schmid, *Large scale image search*, Keynote speaker at the Conference on Machine Vision Applications, Yokohama, Japan, May 2009.
- C. Schmid, *Learning visual human actions from movies*, Seminar at University of Texas at Austin, April 2009.
- C. Schmid, *Burstiness for large scale image search*, Seminar at Oxford University, March 2009.
- C. Schmid, *Learning visual human actions from movies*, Seminar at UCL, London, March 2009.
- C. Schmid, *Large scale image search*, Seminar at ETHZ, Zürich, Feburary 2009.
- J. Verbeek, *Discriminative learning of nearest-neighbor models for image auto-annotation*, Seminar at University of Amsterdam, Intelligent Systems Laboratory, April 2009.
- J. Verbeek, *Improving People Search Using Query Expansions*, Seminar at GREYC Laboratoire, Université de Caen, February 2009.
- J. Verbeek, *Apprentissage semi-supervisé pour la classification d'images*, Colloquium statistiques pour le traitement de l'image, Université Paris 1, Panthéon-Sorbonne, January 2009.
- J. Verbeek, *Machine learning for semantic image interpretation*, seminar Laboratoire Jean Kuntzmann, Grenoble, June 2009.

10. Bibliography

Year Publications

Doctoral Dissertations and Habilitation Theses

[1] R. RONFARD. *Analyse Automatique de films: des sequences d'images aux sequences d'actions*, Université de Grenoble, December 2009, http://lear.inrialpes.fr/pubs/2009/Ron09a, Habilitation à Diriger des Recherches.

Articles in International Peer-Reviewed Journal

[2] 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", 2010, http://lear.inrialpes.fr/pubs/2010/CLTNJ10/, to appear.

[3] V. FERRARI, F. JURIE, C. SCHMID. *From images to shape models for object detection*, in "International Journal of Computer Vision", 2010, http://lear.inrialpes.fr/pubs/2010/FJS10/, to appear.

[4] M. HEIKKILA, M. PIETIKAINEN, C. SCHMID. *Description of interest regions with local binary patterns*, in "Pattern Recognition", vol. 42, n^{O} 3, March 2009, p. 425–436, http://lear.inrialpes.fr/pubs/2009/HPS09.

[5] H. JÉGOU, M. DOUZE, C. SCHMID. *Improving bag-of-features for large scale image search*, in "International Journal of Computer Vision", 2010, http://lear.inrialpes.fr/pubs/2010/JDS10a/, to appear.

[6] 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", vol. 32, n^{o} 1, January 2010, p. 2–11, http://lear.inrialpes.fr/pubs/2010/JSHV10.

[7] D. LARLUS, F. JURIE. *Latent mixture vocabularies for object categorization and segmentation*, in "Journal of Image and Vision Computing", vol. 27, n^{o} 5, April 2009, p. 523-534, http://lear.inrialpes.fr/pubs/2009/LJ09.

[8] 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", 2010, http://lear.inrialpes.fr/pubs/2010/LVJ10/, to appear.

[9] S. MALINOWSKI, H. JÉGOU, C. GUILLEMOT. *Computation of posterior marginals on aggregated state models for soft source decoding*, in "IEEE Transactions on Communications", vol. 57, n^{o} 4, April 2009, p. 888–892, http://lear.inrialpes.fr/pubs/2009/MJG09.

[10] J. VAN DE WEIJER, C. SCHMID, J. VERBEEK, D. LARLUS. *Learning color names for real-world applications*, in "IEEE Transactions on Image Processing", vol. 18, n^{o} 7, July 2009, p. 1512–1523, http://lear.inrialpes.fr/pubs/2009/VSVL09.

International Peer-Reviewed Conference/Proceedings

[11] M. ALLAN, J. VERBEEK. *Ranking user-annotated images for multiple query terms*, in "British Machine Vision Conference", September 2009, http://lear.inrialpes.fr/pubs/2009/AV09.

[12] M. DOUZE, M. GUILLAUMIN, T. MENSINK, C. SCHMID, J. VERBEEK. *INRIA-LEARs participation to ImageCLEF 2009*, in "Working Notes for the CLEF 2009 Workshop", September 2009, http://lear.inrialpes.fr/pubs/2009/DGMSV09.

[13] M. DOUZE, H. JÉGOU, H. SINGH, L. AMSALEG, C. SCHMID. *Evaluation of GIST descriptors for web-scale image search*, in "International Conference on Image and Video Retrieval", ACM, July 2009, http://lear.inrialpes.fr/pubs/2009/DJSAS09.

[14] A. GAIDON, M. MARSZAŁEK, C. SCHMID. *Mining visual actions from movies*, in "British Machine Vision Conference", September 2009, http://lear.inrialpes.fr/pubs/2009/GMS09.

[15] M. GUILLAUMIN, T. MENSINK, J. VERBEEK, C. SCHMID. *TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation*, in "International Conference on Computer Vision", September 2009, http://lear.inrialpes.fr/pubs/2009/GMVS09.

[16] M. GUILLAUMIN, J. VERBEEK, C. SCHMID. *Is that you? Metric learning approaches for face identification*, in "International Conference on Computer Vision", September 2009, http://lear.inrialpes.fr/pubs/2009/GVS09.

[17] H. HARZALLAH, F. JURIE, C. SCHMID. *Combining efficient object localization and image classification*, in "International Conference on Computer Vision", September 2009, http://lear.inrialpes.fr/pubs/2009/HJS09.

[18] T. JIANG, F. JURIE, C. SCHMID. *Learning shape prior models for object matching*, in "Conference on Computer Vision and Pattern Recognition", June 2009, http://lear.inrialpes.fr/pubs/2009/JJS09.

[19] H. JÉGOU, M. DOUZE, C. SCHMID. *On the burstiness of visual elements*, in "Conference on Computer Vision and Pattern Recognition", June 2009, http://lear.inrialpes.fr/pubs/2009/JDS09a.

[20] H. JÉGOU, M. DOUZE, C. SCHMID. *Packing bag-of-features*, in "International Conference on Computer Vision", September 2009, http://lear.inrialpes.fr/pubs/2009/JDS09b.

[21] M. MARSZAŁEK, I. LAPTEV, C. SCHMID. *Actions in context*, in "Conference on Computer Vision and Pattern Recognition", June 2009, http://lear.inrialpes.fr/pubs/2009/MLS09.

[22] H. SANDHAWALIA, H. JÉGOU. *Searching with expectations*, in "International Conference on Acoustics, Speech, and Signal Processing", Signal Processing, March 2010, http://lear.inrialpes.fr/pubs/2010/SJ10, to appear.

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

[24] H. WANG, M. M. ULLAH, A. KLÄSER, I. LAPTEV, C. SCHMID. *Evaluation of local spatio-temporal features for action recognition*, in "British Machine Vision Conference", September 2009, http://lear.inrialpes.fr/pubs/2009/WUKLS09.

National Peer-Reviewed Conference/Proceedings

[25] M. GUILLAUMIN, T. MENSINK, J. VERBEEK, C. SCHMID. *Apprentissage de distance pour l'annotation d'images par plus proches voisins*, in "Reconnaissance des Formes et Intelligence Artificielle", January 2010, http://lear.inrialpes.fr/pubs/2010/GMVS10a, to appear.

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

Scientific Books (or Scientific Book chapters)

[27] H. JÉGOU, M. DOUZE, C. SCHMID. *Recent advance in image search*, in "Emerging Trends in Visual Computing", F. NIELSEN (editor), Springer, 2009, p. 305–326, http://lear.inrialpes.fr/pubs/2009/JDS09.

[28] S. LAZEBNIK, C. SCHMID, J. PONCE. *Spatial pyramid matching*, in "Object Categorization: Computer and Human Vision Perspectives", S. DICKINSON, A. LEONARDIS, B. SCHIELE, M. TARR (editors), chap. 21, Cambridge University Press, 2009, p. 401–415, http://lear.inrialpes.fr/pubs/2009/LSP09.

Research Reports

[29] H. JÉGOU, M. DOUZE, C. SCHMID. *Searching with quantization: approximate nearest neighbor search using short codes and distance estimators*, n^{o} RR-7020, INRIA, August 2009, http://lear.inrialpes.fr/pubs/2009/JDS09d, Technical report.