INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team LEAR

Learning and Recognition in Vision

Grenoble - Rhône-Alpes

THEME COG

Activity Report

2008

Table of contents

1. Team **1**

2. Overall Objectives **1**

2.1. Introduction 1

2.2. Highlights of the year 2

3. Scientific Foundations **3**

3.1. Image features and descriptors and robust correspondence 3

3.2. Statistical modeling and machine learning for image analysis 4

3.3. Visual recognition and content analysis 4

4. Application Domains **5**

5. Software **6**

5.1. Large-scale image indexing 6

5.2. Extracting and describing interest points 6

5.3. Image search demonstrator 6

5.4. Datasets 6

6. New Results **8**

6.1. Large-scale image search 8

6.1.1. Hamming Embedding 8

6.1.2. Weak geometry consistency 8

6.1.3. Large-scale video search 9

6.1.4. Approximate nearest neighbor search 9

6.2. Semi-supervised learning and structuring of visual models 10

6.2.1. Automatic face naming from caption-based supervision 10

6.2.2. Improving people search using query expansion 11

6.2.3. Learning metrics for visual identification 11

6.2.4. Learning image categories from web data 12

6.2.5. Semi-supervised dimensionality reduction using pairwise equivalence constraints 15

6.2.6. Constructing category hierarchies for visual recognition 15

6.3. Supervised visual object recognition 15

6.3.1. Learning shape prior models for object matching 15

6.3.2. Viewpoint-independent object class detection using 3D feature maps 16

6.3.3. Object recognition by integrating multiple image segmentations 17

6.3.4. Category-level object segmentation 17

6.3.5. Classification aided two step localization 18

6.4. Video interpretation 18

6.4.1. Learning realistic human actions from movies 18

6.4.2. Actions in context 18

6.4.3. Spatio-temporal descriptor for action recognition 20

6.4.4. Action recognition in video sequences with and without human localization 21

7. Contracts and Grants with Industry **21**

7.1. Start-up Milpix 21

7.2. MDBD Aerospatiale 22

7.3. MSR-INRIA joint lab: scientific image and video mining 22

8. Other Grants and Activities **22**

8.1. National Projects 22

8.1.1. QUAERO 22

8.1.2. ANR Project GAIA 22

8.1.3. ANR Project RAFFUT 22

8.1.4. ANR Project R2I 23

8.1.5. ANR Project RobM@rket 23

8.1.6. GRAVIT Grant 23
8.2. International Projects 23
8.2.1. FP6 European Project CLASS 23
8.2.2. FP7 European Network of Excellence PASCAL 2 24
8.2.3. FP6 Marie Curie EST Host Grant VISITOR 24
8.3. Bilateral relationships 24
9. Dissemination 24
9.1. Leadership within the scientific community 24
9.2. Teaching 25
9.3. Invited presentations 25
10. Bibliography 26

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 [INRIA Research Director, project-team leader, HdR]
Hervé Jégou [INRIA Researcher]
Jakob Verbeek [INRIA Researcher]

Faculty Member

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

External Collaborator

Frédéric Jurie [Professor at University of Caen]

Technical Staff

Matthijs Douze [ANR project RAFFUT until 07/08, QUAERO project 08/08–12/10]
Yves Gufflet [EU project CLASS, 05/06–05/08]
Benoit Mordelet [GRAVIT project 09/07–08/08, ANR project RAFFUT 09/08–12/09]
Christophe Smekens [INRIA, ODL, 09/07–02/09]

PhD Student

Adrien Gaidon [INPG, Microsoft/INRIA project, from 10/08]
Matthieu Guillaumin [INPG, Ministry of research grant from 09/07]
Hedi Harzallah [INPG, MBDA project, from 02/07]
Alexander Kläser [INPG, EU project CLASS, from 11/06]
Josip Krapac [University of Caen, ANR project R2I, from 01/08]
Diane Larlus [INPG, Ministry of research grant, 10/05–11/08]
Joerg Liebelt [INPG, EADS scholarship, co-supervised with TU Munich, from 10/06]
Marcin Marszalek [INPG, Marie Curie project VISITOR, 09/05–12/08]

Post-Doctoral Fellow

Moray Allan [EU project CLASS, 12/07–12/09]
Tingting Jiang [INRIA, 12/07–05/09]

Visiting Scientist

Krystian Mikolajczyk [University of Surrey, regular visits]

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:

- **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 TRECVID copyright detection task & PASCAL VOC'2008 challenge**. LEAR participated in the copyright detection task of the TRECVID'2008 evaluation campaign organized by the National Institute for Standard and Technologies. We obtained excellent results, i.e., top accuracy for all types of transformations (camcording, compression, etc). See http://www-nlpir.nist.gov/projects/trecvid for more information.

  LEAR also obtained excellent results in the PASCAL VOC'2008 detection and classification tasks. LEAR achieved best detection and classification results for 11 out of 20 classes and 8 out of 20 classes, see http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008.
- **Platform and on-line demo for fast image search in large databases**. LEAR has developed an image indexing platform that searches in real time for similar images in very large databases. This platform can retrieve corresponding images even if the query image has undergone significant changes, as for example an important scale change. LEAR is currently transferring and testing this platform to the start-up MILPIX, created by a former LEAR PhD student. A public on-line demo can be accessed from the LEAR homepage at http://lear.inrialpes.fr.
- **International Workshop on Object Recognition.** LEAR co-organized the 4th International Workshop on Object Recognition in May 2008. The objective of this three day workshop was to discuss and advance the state-of-the-art in visual object recognition in images and videos. Many key players

of the field attended the workshop. The workshop was co-organized by A. Efros (CMU), C. Schmid (LEAR), A. Torralba (MIT) and T. Tuytelaars (KU Leuven). It received funding from Adobe, Imra, Inria, Intel, Honda, Kitware, Microsoft, Mitsubishi, Toyota and Xerox. See http://lear.inrialpes.fr/RecogWorkshop08 for more information.

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.

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/LLA.
- 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-learning 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, image 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 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 even 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 investigate 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.

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 and the changeable appearance of objects due to global and local temperature changes. Our industrial project with MBDA is on detecting objects, for example cars, observed from airplanes or missiles. Additional difficulties are the presence of severe changes of the imaging conditions and the small size of object regions.

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

5. Software

5.1. Large-scale image indexing

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

LEAR has developed an image search engine named BIGIMBAZ. This software queries for similar images in a very large database: currently two million images in a few seconds on a single core processor. This platform integrates several of LEAR's scientific contributions on large-scale indexing: retrieval using our Hamming Embedding method and applying partial geometrical information on a large scale [22].

The image search engine is under copyright protection, i.e.,

- the first version was registered at the APP (Agence pour la Protection des Programmes) under the identifier IDDN.FR.001.130028.000.S.P.2007.000.10300.
- the second version is registered at the APP (IDDN.FR.001.510004.000.S.P.2008.000.21000).
- the technology is protected by a INRIA patent (Ptaent application no. 08/03345, filed on June 16th 2008). The patent covers the technology of the second version of the search engine.

The image search engine has been licensed to the start-up MILPIX, in charge of its commercial exploitation.

5.2. Extracting and describing interest points

Participants: Matthijs Douze, Hervé Jégou, Cordelia Schmid, Christophe Smekens.

Local descriptors [41] computed at affine invariant local regions [42] provide a stable image characterization in the presence of significant viewpoint changes. They allow for robust image correspondence despite large changes in viewing conditions, which in turn allows rapid appearance based indexing in large image databases. Over the past several years we have been developing efficient software for this, see http://lear.inrialpes.fr/software.

This year, LEAR has developed a new library for local description, called OBSIDIAN. It has been created in order to meet the requirements of industrial large-scale applications and to prevent intellectual property issues. The library integrates robust detection of regions of interest and the CSLBP descriptor [9]. It is used, in particular, in conjunction with the image search engine BIGIMBAZ.

The software was registered at the APP under the identifier IDDN.FR.001.280044.000.S.P.2008.000.10300. It has been licensed to the start-up MILPIX.

5.3. Image search demonstrator

Participants: Matthijs Douze, Hervé Jégou, Cordelia Schmid, Christophe Smekens.

Building upon our core image system BIGIMBAZ and our description library OBSIDIAN, we have designed a public on-line image search demonstrator. This demonstrator searches an image or a part of it among two million images. The on-line demonstrator is accessible on the LEAR webpage, http://lear.inrialpes.fr. Figure 1 shows the web interface and the results obtained for a query.

5.4. Datasets

Participants: Matthijs Douze, Hervé Jégou, Alexander Kläser, Ivan Laptev [Vista, INRIA Rennes], Marcin Marszalek, Cordelia Schmid.

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, i.e., include occlusions, viewpoint changes, illumination changes, etc. Benchmarking procedures allow to compare the strengths of different approaches. Providing a clear and broadly understood performance measure is, therefore, essential.

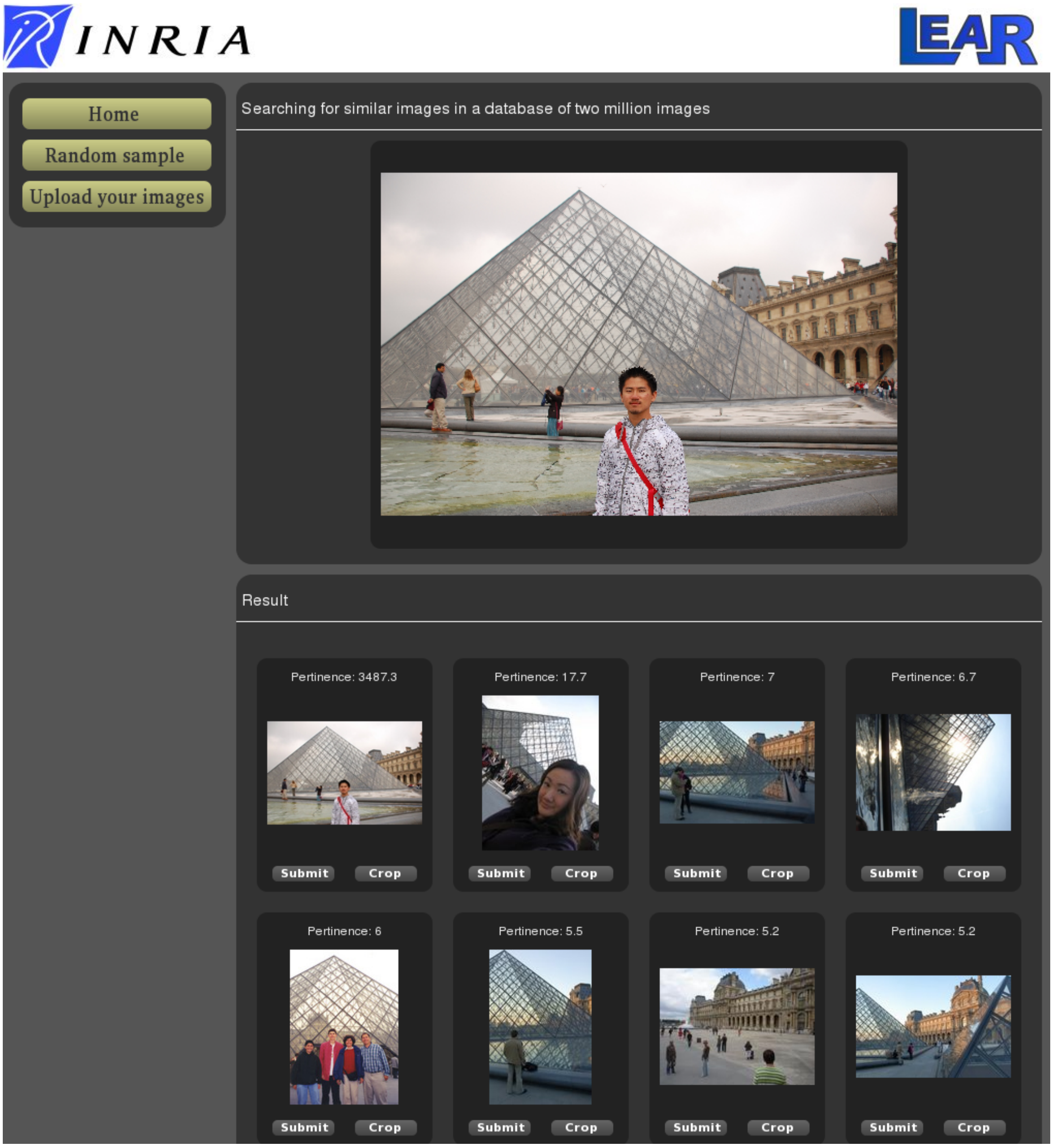


Figure 1. LEAR's on-line image search demonstrator. The top image is the query and the remaining images are the top results retrieved from a database of two (now ten) million images (ordered from left to right line by line). Note that the query image itself is contained in the database and returned as the strongest response. The demonstrator can be tested at http://lear.inrialpes.fr.

In addition to the datasets previously created by the project-team, we have designed two new datasets this year, INRIA Holidays and Hollywood Human Action. There are described in the following. All our datasets are accessible and can be downloaded at http://lear.inrialpes.fr/data.

INRIA Holidays dataset:

We have designed a new image dataset in order to evaluate the relevance of our image search algorithms for personal photos. This dataset, called INRIA Holidays, has been created by collecting personal holiday photos containing groups of photos representing the same building or scene. The remaining ones were taken on purpose to test the robustness to various transformations: rotations, viewpoint and illumination changes, blurring, etc. The dataset includes a very large variety of scene types (natural, man-made, water and fire effects, etc) and images are of high resolution. The dataset contains 500 image groups, each of which represents a distinct scene, for example San Marco Square in Venice. The first image of each group is the query image and the correct retrieval results are the other images of the group. This dataset can be downloaded at http://lear.inrialpes.fr/people/jegou/data.php.

In order to assess the efficiency of our algorithms on a large scale, we have also downloaded 10 million distractor images from the web.

Hollywood Human Actions dataset:

The Hollywood Human Actions dataset contains video samples with human actions from 32 movies. Each sample is labeled according to one or more of 8 action classes: AnswerPhone, GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp, StandUp. The dataset is divided into a test set obtained from 20 movies and two training sets obtained from 12 movies different from those of the test set. The automatic training set is obtained using automatic script-based action annotation and contains 233 video samples with approximately 60% correct labels. The clean training set contains 219 video samples with manually verified labels. The test set contains 211 samples with manually verified labels. The dataset was originally used in [25] and can be downloaded at http://lear.inrialpes.fr/data.

6. New Results

6.1. Large-scale image search

6.1.1. Hamming Embedding

Participants: Matthijs Douze, Hervé Jégou, Cordelia Schmid.

The bag-of-features representation proposed by Sivic and Zisserman [43] is an efficient and accurate way of representing images. It is the starting point of several recent state-of-the-art contributions on large scale image search. However, the representation underlying this approach fails to properly exploit the discriminative power provided by state-of-the-art local descriptors due to quantization. To address this issue, we have proposed a finer representation of local descriptors. The method used, called Hamming Embedding [22], can be seen as an extension of bag-of-features representation. The key idea is to refine the descriptor representation by adding a short binary signature that provides a better localization of the descriptor within the quantizing cell, as illustrated by Figure 2. The binary signature is generated by an Euclidean-to-Hamming mapping function, designed such that the neighborhood of a descriptor for the Euclidean distance is approximately represented by the neighborhood of the binary signature in the Hamming space. As a result, Hamming Embedding provides a state-of-the-art approximate nearest neighbor search algorithm.

6.1.2. Weak geometry consistency

Participants: Matthijs Douze, Hervé Jégou, Cordelia Schmid.

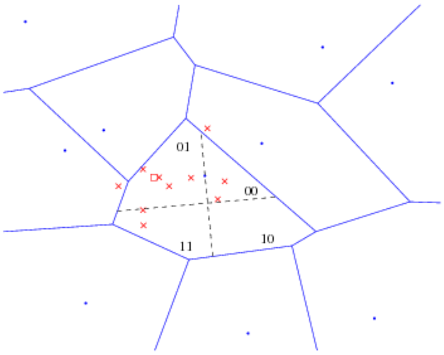


Figure 2. Illustration of k-means clustering and our binary signature: the similarity search is based on the Voronoi cell index and the Hamming distance between binary signatures. We display the noisy versions, $\times$, of a given descriptor $\square$: most of the noisy data points fall into bins with a small Hamming distance to the descriptor, i.e., there is no point in the cell 10.

A drawback of the bag-of-features representation is that is does not take into account any geometric information. Geometric consistency is in general exploited in a final re-ranking stage by estimating an affine transformation mapping the query image onto the database image. However, the amount of database images for which this step is performed is limited by the complexity of such an estimation. This limitation is problematic for large datasets, because in this case the short-list has a length which is negligible compared to the database size, and relevant images might not be ranked well enough to be considered in the re-ranking stage.

The key idea of weak geometry consistency [22] is to use only partial geometric information, but for all the database images in a very efficient manner. This is done by integrating geometric constraints within the inverted file system. To be more precise we exploit the consistency of scale and orientation differences between images. As a result, the search penalizes the descriptors that are not consistent in terms of orientation angle and scale, as illustrated in Figure 3.

6.1.3. Large-scale video search

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

We have recently addressed [18] the problem of large-scale video search, i.e., retrieval of the corresponding video from a large collection of videos in the presence of severe image transformations and temporal crops. The indexing system we propose is derived from our image search system: it uses local descriptors extracted from image regions, the Hamming Embedding method detailed in 6.1.1 as well as the weak geometry constraints of 6.1.2 to index and search the frames of a given video. Temporal information is integrated based on outlier removal with a Hough transform.

Our video search engine has been used for the copy detection task of the TRECVID'2008 evaluation campaign, where 22 participants have submitted runs. We obtained top results in terms of accuracy for all types of transformations, see [18] for details.

6.1.4. Approximate nearest neighbor search

Participants: Laurent Amsaleg [TexMex, INRIA Rennes], Patrick Gros [TexMex, INRIA Rennes], Hervé Jégou, Cordelia Schmid.

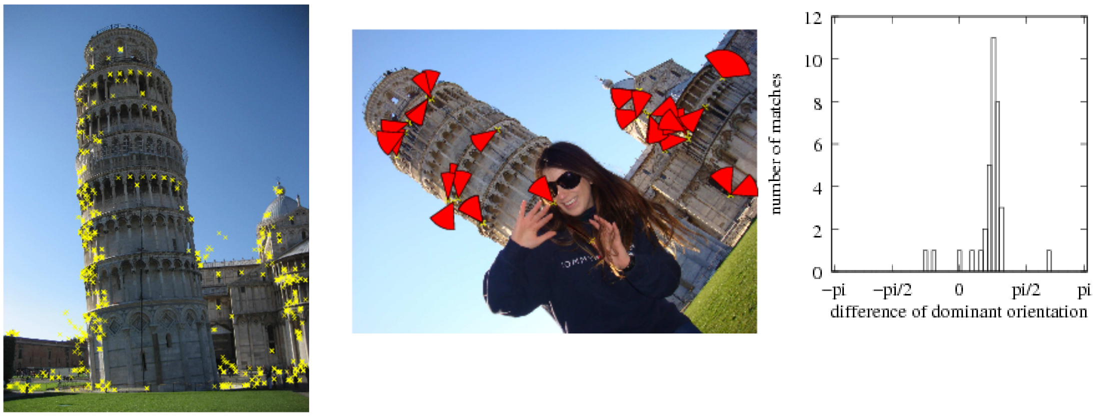


Figure 3. Orientation consistency. Left: Query image and its interest points. Middle: An image of the same location viewed under a different image rotation. The slices show the differences between the estimated dominant orientations of the query interest points and the image interest points. Right: Corresponding histogram of dominant orientation differences: The peak corresponds to the global angle variation. The outliers are filtered.

It is well known that high-dimensional nearest neighbor retrieval is very expensive and that image indexation methods suffer from its computational cost. A significant performance gain can be obtained by using approximate nearest neighbor search methods, as for example Locality-Sensitive Hashing (LSH). In [21] we have proposed an improvement of LSH which performs on-line selection of the most appropriate hash functions from a pool of functions. For this purpose, we have shown that a measure of the expected relevance of a given hash function can be computed without parsing the corresponding hashing cells involved in LSH. An additional improvement originates from the use of E_8 lattices for geometric hashing instead of one-dimensional random projections.

6.2. Semi-supervised learning and structuring of visual models

6.2.1. Automatic face naming from caption-based supervision

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

Over the last decades large digital multimedia archives have appeared, through digitization efforts by broadcasting services, through news oriented media publishing on-line, or through user provided content concentrated on websites such as YouTube and Flickr. There is a broad ongoing effort to develop methods to disclose such archives in a user oriented and semantically meaningful way. In particular, there is a great interest in 'unsupervised' systems for automatic content analysis in such archives, that do not require manual annotations to link content to semantic concepts. We are, in particular, interested in finding images of people on the web, and more specifically within databases of captioned news images. The goal is to automatically retrieve faces of a person.

We consider two scenarios of naming people in databases of news photos with captions: (i) finding faces of a single person, and (ii) assigning names to all faces. We have combined an initial text-based step, that restricts the name assigned to a face to the set of names appearing in the caption, with a second step that analyzes visual features of faces. By searching for groups of highly similar faces that can be associated with a name, the results of purely text-based search can be greatly improved [20].

We built upon a recent graph-based approach in which nodes correspond to faces and edges connect highly similar faces. We improved this approach by introducing constraints when optimizing the objective function that limit the method to return at most one face for each name per image, and propose improvements in the low-level methods used to construct the graphs. Furthermore, we have generalized the graph-based approach to face naming in the full data set. In this multi-person naming case the optimization quickly becomes computationally demanding, and we have developed an important speed-up using graph-flows to compute the optimal name assignments in documents. Generative models had previously been proposed to solve the multi-person naming task. We have compared the generative and graph-based methods in both scenarios, and obtain significantly better performance using the graph-based method in both cases.

6.2.2. Improving people search using query expansion

Participants: Thomas Mensink, Jakob Verbeek.

The underlying idea of the approach described in the previous paragraph is that text-based search will render the queried person to be relatively frequent as compared to other people, so we can search for a large group of highly similar faces. The performance depends strongly on this assumption: for people whose face appears in less than about 40% of the initial text-based result, the performance may be very poor. Our contribution [29] is to improve search results by exploiting faces of other people that co-occur frequently with the queried person. We refer to this process as ‘query expansion’.

We use the query expansion to provide a query-specific relevant set of ‘negative’ examples which should be separated from the potentially positive examples in the text-based result set. We apply this idea to a recently proposed method which filters the initial result set using a Gaussian mixture model, and apply the same idea using a logistic discriminant model. We experimentally evaluate the methods using a set of 23 queries on a database of 15.000 captioned news stories from Yahoo news.

Our experimental results show that query expansion leads to improved results when searching for people in captioned news images. Although queries for which text-based search in the caption leads to a low fraction of relevant faces remain difficult, we have made significant progress in these cases, boosting precision by 20% up to 50% (absolute increase) for the generative model in the five most difficult cases. We achieve performance levels that are significantly higher than the state-of-the-art. We obtain our results with methods that do not require calculation of pairwise similarities, which are therefore fast when many faces are processed. Our best method (logistic discriminant + query expansion) obtains a precision of 84% for a recall of 85%, while the best previously reported result on these queries only reaches a precision of 73% for the same recall of 85%. For reference, when simply returning all faces that were detected in images with the queried name in the caption, a precision of 44% is obtained.

A demonstration of the generative model with query expansion on the Yahoo news data set is available at http://lear.inrialpes.fr/~verbeek/facefinder. See figure 4 for a few examples.

6.2.3. Learning metrics for visual identification

Participants: Matthieu Guillaumin, Cordelia Schmid, Jakob Verbeek.

Visual identification is a binary classification problem, in which we have to decide whether two images depict the same object or not. We have considered several discriminative approaches for visual identification: (i) learning a metric between the image representations using a simple logistic discriminant framework, and (ii) using a k-nearest neighbour classifier to estimate the marginal probability that the two images belong to the same class.

Distances that are linear with respect to their parameters can be optimized in a standard logistic discriminant model with maximum likelihood estimation. We apply this idea to the Mahalanobis distance between features extracted from two images. By specifying, in a supervised framework, which image pairs depict the same object and which don’t, the model is learned to split positive and negative image pairs, and provides the optimal threshold on the distance for classification purposes. The resulting distance (LDPM) is a pseudo-metric, it is very fast to learn and evaluate, and outperforms the existing state-of-the-art.

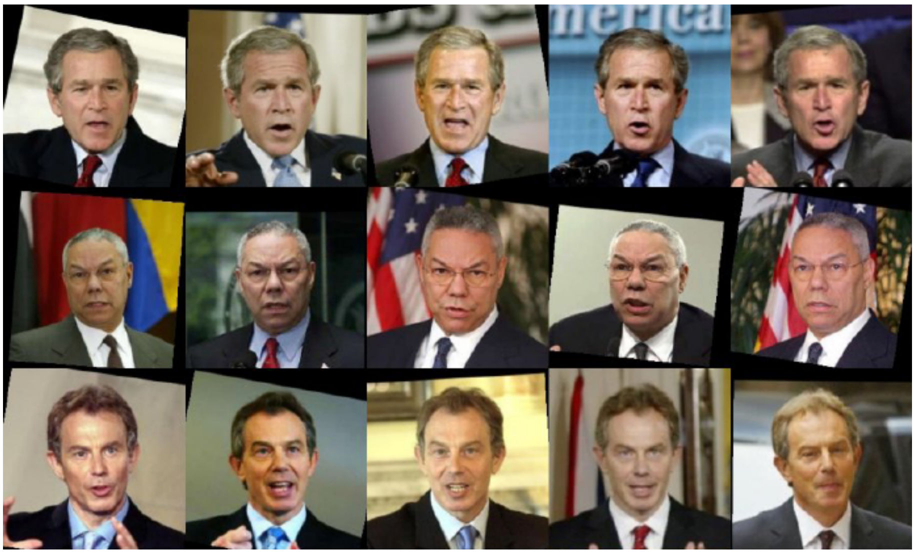


Figure 4. Example faces that are retrieved on the basis of a text query for the name of three persons. Note that these faces are retrieved without ever using any manual annotation, all association between the query name and the faces is automatically generated from the captions that come with the images.

The second approach assigns scores to pairs by implicitly using all positive and negative pairs that can be generated from the labelled data, and not just labelled pairs as above. Using a base metric, we recover the k-nearest neighbourhoods of both images, and count how many positive pairs can be made from these neighbourhoods. This score can be understood as the probability that the two images belong to the same class, marginalised over classes. This approach yields very good results. The drawback of the method is the computation cost of finding the nearest neighbours, which makes it unrealistic for large-scale use without resorting to approximate techniques.

The learned LDPM metric being very fast to evaluate, we can use it to improve results for applications like recognition from a single exemplar, for which we show a significant improvement over L2 results. We also applied our learned metric to the problem of unsupervised clustering of faces, see figure 5 for two examples clusters. Again, a very significant improvement over L2 is obtained.

6.2.4. Learning image categories from web data

Participants: Moray Allan, Cordelia Schmid, Jakob Verbeek.

When searching for images from the web based on keywords, the resulting set is “noisy”, i.e., only some of the images correspond visually to the keyword. Given such a collection of training images partially labelled with ‘noisy’ annotations, we wish to learn image and scene categories. The learnt category models can be used to find relevant database images in response to user search queries, or to answer questions about specific images. We are currently working with a dataset of images downloaded from the Flickr photo sharing website. Each image may have a number of ‘tags’ associated with it, describing aspects of the image. These tags may be, for example, the names of people and locations that feature in an image, but often correspond to more general categories. We use the tag information together with visual features from the images to learn object category models, providing cleaner search results than are achieved by using tags or visual features alone. Our current focus is on improving the retrieval accuracy for queries with multiple search terms, as illustrated by Fig. 6.

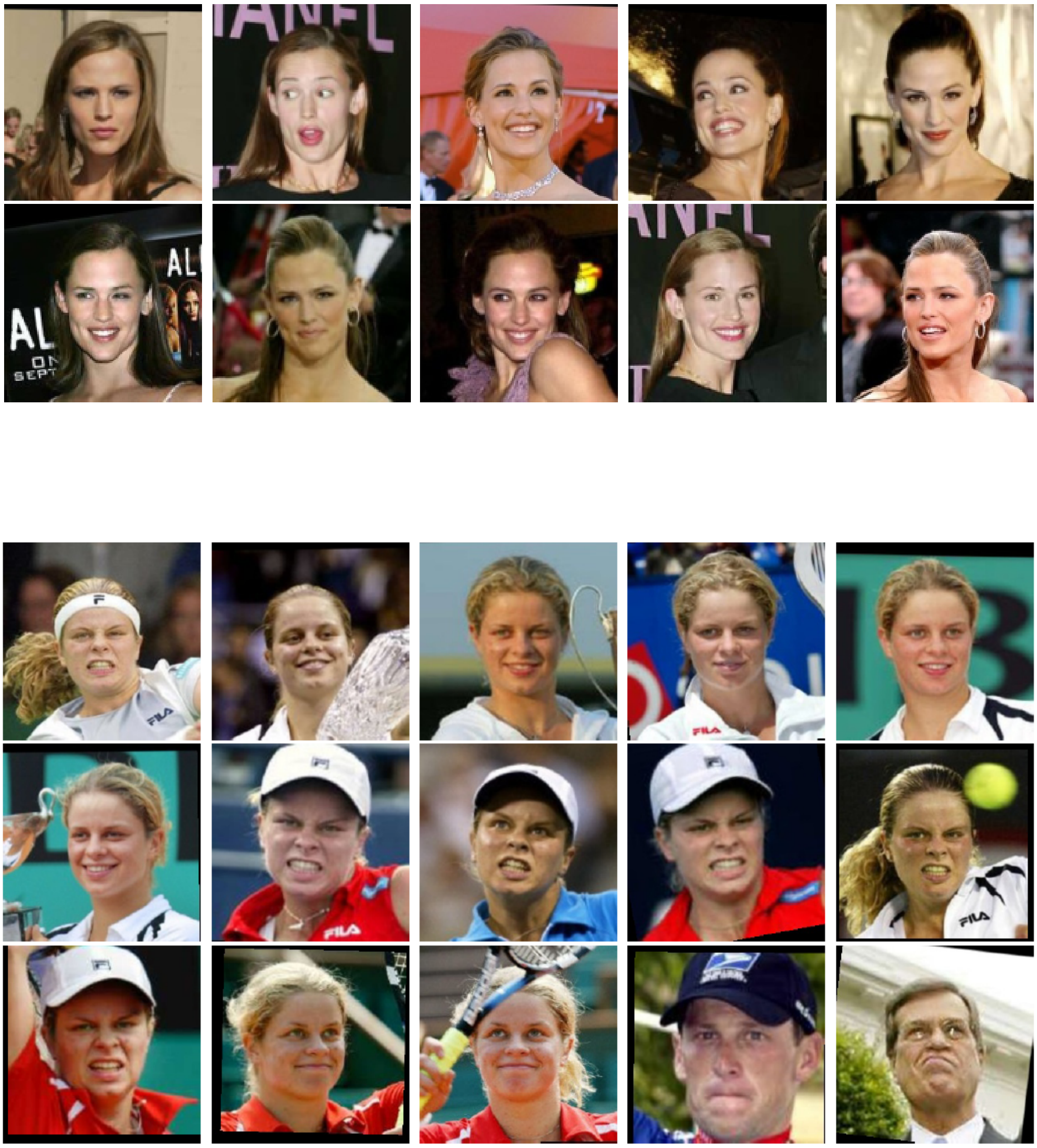


Figure 5. Two example clusters obtained using the LDPM metric. The top cluster is pure, only 2 faces of this person were assigned to other clusters. The bottom cluster is more typical: it contains a few faces of other people (the last 2 images).

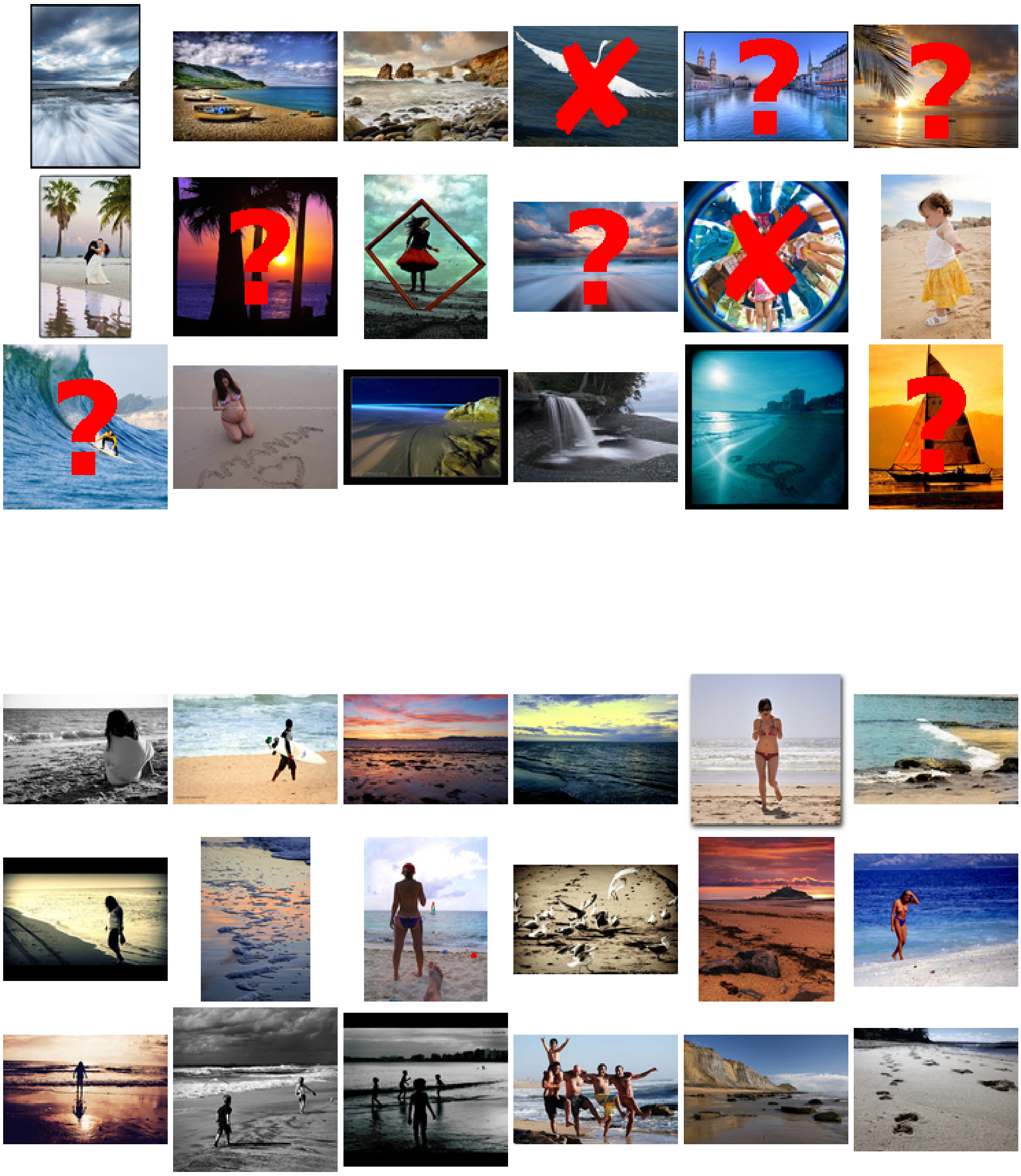


Figure 6. Top: First 18 images returned by Flickr for 'beach'. Bottom: Flickr 'beach' images after visual feature based re-ranking. Crosses indicate the incorrect images and question marks the doubtful ones.

6.2.5. Semi-supervised dimensionality reduction using pairwise equivalence constraints

Participants: Hakan Cevikalp, Frédéric Jurie, Alexander Kläser, Jakob Verbeek.

The standard way in which classification problems are solved is to collect a set of input patterns labeled with the correct class, and then to fit a classification function on these input-output pairs. However, in many applications obtaining labels is a costly procedure as it often requires human effort. On the other hand, in some applications side information –given in the form of pairwise constraints indicating that two input patterns belong to the same class or to different classes– is available without or with little extra cost. For instance, faces extracted from successive video frames in roughly the same location can be assumed to represent the same person, whereas faces extracted in different locations in the same frame can be assumed to be from different persons.

In [17] we propose a semi-supervised dimensionality reduction scheme which uses side information in form of pairwise equivalence constraints to improve clustering and classification performance. Our algorithm first finds neighboring points for each input to create a weighted neighborhood graph. Then, the side-information constraints are used to modify the neighborhood relations and weight matrix to reflect this weak form of supervision. The optimal projection matrix according to our cost function is then identified by solving for the smallest eigenvalue solutions of a $n \times n$ eigenvector problem, where n is the number of input patterns. Experimental results show that our semi-supervised dimensionality reduction method increases performance of subsequent clustering and classification algorithms.

6.2.6. Constructing category hierarchies for visual recognition

Participants: Marcin Marszalek, Cordelia Schmid.

Class hierarchies are commonly used to reduce the complexity of the classification problem. This is crucial when dealing with a large number of categories. In [28] we evaluate class hierarchies currently constructed for visual recognition. We show that top-down as well as bottom-up approaches, which are commonly used to automatically construct hierarchies, incorporate assumptions about the separability of classes. Those assumptions do not hold for visual recognition of a large number of object categories. We therefore propose a modification which is appropriate for most top-down approaches [28]. It allows to construct class hierarchies that postpone decisions in the presence of uncertainty and thus provide higher recognition accuracy. We also compare our method to a one-against-all approach and show how to control the speed-for-accuracy trade-off with our method. For the experimental evaluation, we use the Caltech-256 visual object classes dataset and compare to state-of-the-art methods. Experimental results show that the constructed hierarchies are similar to hierarchies extracted from semantic networks, even though the hierarchy is based on visual data only. Unlike the purely semantic hierarchies, however, it also groups classes that are related by semantic links difficult to model, or that feature accidental similarity.

We also propose to automatically extract vision-oriented semantic information from Flickr. Building on image tags - semantics provided by Flickr users - we show how to construct a rich class hierarchy that reflects visual similarities between classes. In our automatically built class hierarchies we observe semantic relationships similar to the ones present in expert ontologies, but we also discover visual context links and scene-type groupings. Our experiments show the improved performance of our vision-oriented hierarchies over ontology-based hierarchies in terms of modeling the latent visual similarities between object classes.

6.3. Supervised visual object recognition

6.3.1. Learning shape prior models for object matching

Participants: Tingting Jiang, Frédéric Jurie, Cordelia Schmid.

The aim of this work is to learn a shape prior model for an object class and to improve shape matching with the learned shape prior. Given images of example instances, we can learn a mean shape of the object class as well as the variations of non-affine and affine transformations separately based on the thin plate spline (TPS) parameterization. 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 test our approach by using it to learn shape prior models for all five object classes in the ETHZ shape dataset. The results show that the learning accuracy is better than previous work and the learned shape prior models are helpful for object matching in real applications such as object classification.

Figure 7 shows the mean shapes learned from 5 different training sets for each of the 5 object classes in the ETHZ dataset. Figure 8 (a) is the matching result with the original TPS-RPM [39] using the model points learned by our approach. Figure 8 (b) displays the matching result with the constrained shape matching method by [40]. Figure 8 (c) shows our matching result which is robust to the clutter.

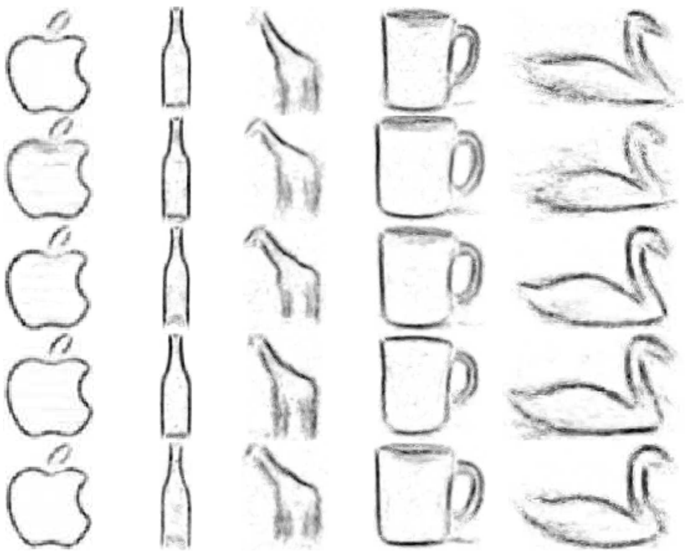


Figure 7. Learned mean shapes from 5 different training sets for each of the 5 object classes in the ETHZ dataset.

6.3.2. Viewpoint-independent object class detection using 3D feature maps

Participants: Joerg Liebelt, Cordelia Schmid.

Most existing approaches to viewpoint-independent object class detection combine classifiers for a few discrete views. We propose instead to build 3D representations of object classes which allow to handle viewpoint changes *and* intra-class variability [27]. We do not build a model from 2D features and their geometric constraints. Instead, we resort to a database of existing, fully textured synthetic 3D models. Our approach renders the synthetic models from different viewpoints and extracts a set of pose and class discriminative features. Discriminative features are obtained by a filtering procedure which identifies features suitable for reliable matching to real image data. A codebook is created based on clusters of these synthetic features encoded by their appearance and 3D position. During detection local features from real images are matched to the synthetically trained ones as part of a probabilistic voting scheme. Each match casts votes to

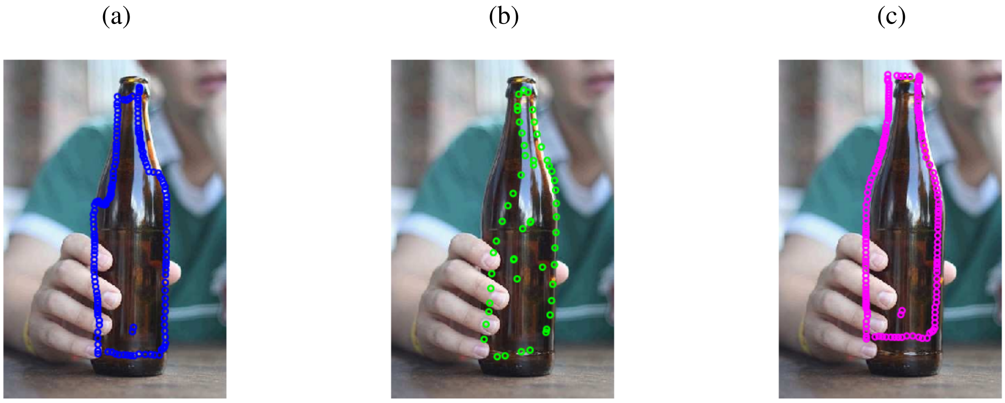


Figure 8. Comparison of different shape matching methods based on the same initialization. (a) Output shape with original TPS-RPM [39] using the mean shape learned by our method. (b) Output shape with the constrained TPS-RPM [40] using their learned shape model. (c) Output shape with our shape matching method using the shape prior learned by our method. The shape models used by [40] and by our method are learned from the same training data.

determine the most likely class and 3D pose of the detected generic object. The most promising votes are then evaluated and refined based on a robust pose estimation step which outputs a 3D bounding box in addition to the 2D localization. On a set of calibrated images, we could show that the estimated 3D pose is sufficient to initialize 3D tracking and model registration. On the PASCAL 2006 dataset for motorbikes and cars, the 2D localization results of our approach competes with state-of-the-art 2D object detectors.

6.3.3. Object recognition by integrating multiple image segmentations

Participants: Martial Hebert [CMU], Caroline Pantofaru [CMU], Cordelia Schmid.

The joint tasks of object recognition and segmentation from a single image are complex in their requirement of not only correct classification, but also deciding exactly which pixels belong to the object. Iterating through all possible pixel subsets is prohibitively expensive, leading to recent approaches which use bottom-up unsupervised image segmentation to reduce the size of the configuration space and create better spatial support for features. Bottom-up image segmentation, however, is known to be unstable, with small image perturbations, feature choices, or different segmentation algorithms leading to drastically different image segmentations. This instability has led to advocacy for using multiple segmentations of an image, and makes the method of combining information from these segmentations crucial.

Our approach [30] explores the question of how to best integrate the information from multiple bottom-up segmentations which are created using different scales, algorithms and features. We show how the use of all the bottom-up segmentations in concert leads to improved object segmentation accuracy and creates robustness to outlier image segmentations. Our intuitive formulation shows performance comparable or better than the state-of-the-art on two difficult datasets, the MSRC 21-class dataset and the PASCAL Visual Object Challenge 2007 segmentation dataset.

This is joint work with Caroline Pantofaru (CMU) and Martial Hebert (CMU); it was partially funded by our associated team Tethys (see 8.3.1).

6.3.4. Category-level object segmentation

Participants: Frédéric Jurie, Diane Larlus, Eric Nowak, Jakob Verbeek.

We have proposed an approach [26], [33], [38] for segmenting objects of a given category, where the category is defined by a set of training images. This problem is also known as *figure-ground segmentation*. Our method is based on a local classification of the patches and a model of their spatial relations. The local classification relies on a bag-of-words representation and combines local patches into larger regions, each representing an object instance or the background. The number of regions is automatically chosen for each test image. Spatial relations are enforced by regularizing the assignment of patches to regions based on a Markov Random Field. This allows to obtain clean boundaries and to enforce label consistency, guided by local image cues (color, texture and edge cues) and by long-distance dependencies. Gibbs sampling is used to infer the model. Our method successfully segments object categories with highly varying appearance, in the presence of cluttered backgrounds and large viewpoint changes. It is able to differentiate between different instances of the same category. We show that it outperforms published segmentation results on the PASCAL VOC'2007 challenge.

6.3.5. Classification aided two step localization

Participants: Hedi Harzallah, Frédéric Jurie, Cordelia Schmid.

Our approach for localizing object categories with a bounding box is based on a sliding window approach, i.e., a classifier decides for a set of multi-scale windows extracted at multiple positions and scales in an image if the target object is present or not. We use a two step classifier: the first one rapidly rejects most of the windows not containing the object; the second one performs a more expensive evaluation on the remaining ones. In step one, we use a linear Support Vector Machine (SVM) classifier and simple histogram of gradient features to filter most of the negative examples. In step two, we apply a non-linear SVM classifier with a χ^2-kernel and combine different types of features. The features used are SIFT features represented by a spatial pyramid as well as histogram of gradient features. Experimental results show that the results of a non-linear classifier significantly improve over a linear one, and that the filtering step reduces the average precision only insignificantly.

Furthermore, we use context hints by combining an image classification score with the object localization score to obtain the final results. This approach won the detection contest of the PASCAL VOC Challenge 2008 for 11 out of 20 categories, see http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008. See Figure 9 for examples of correct, false and missed detections.

6.4. Video interpretation

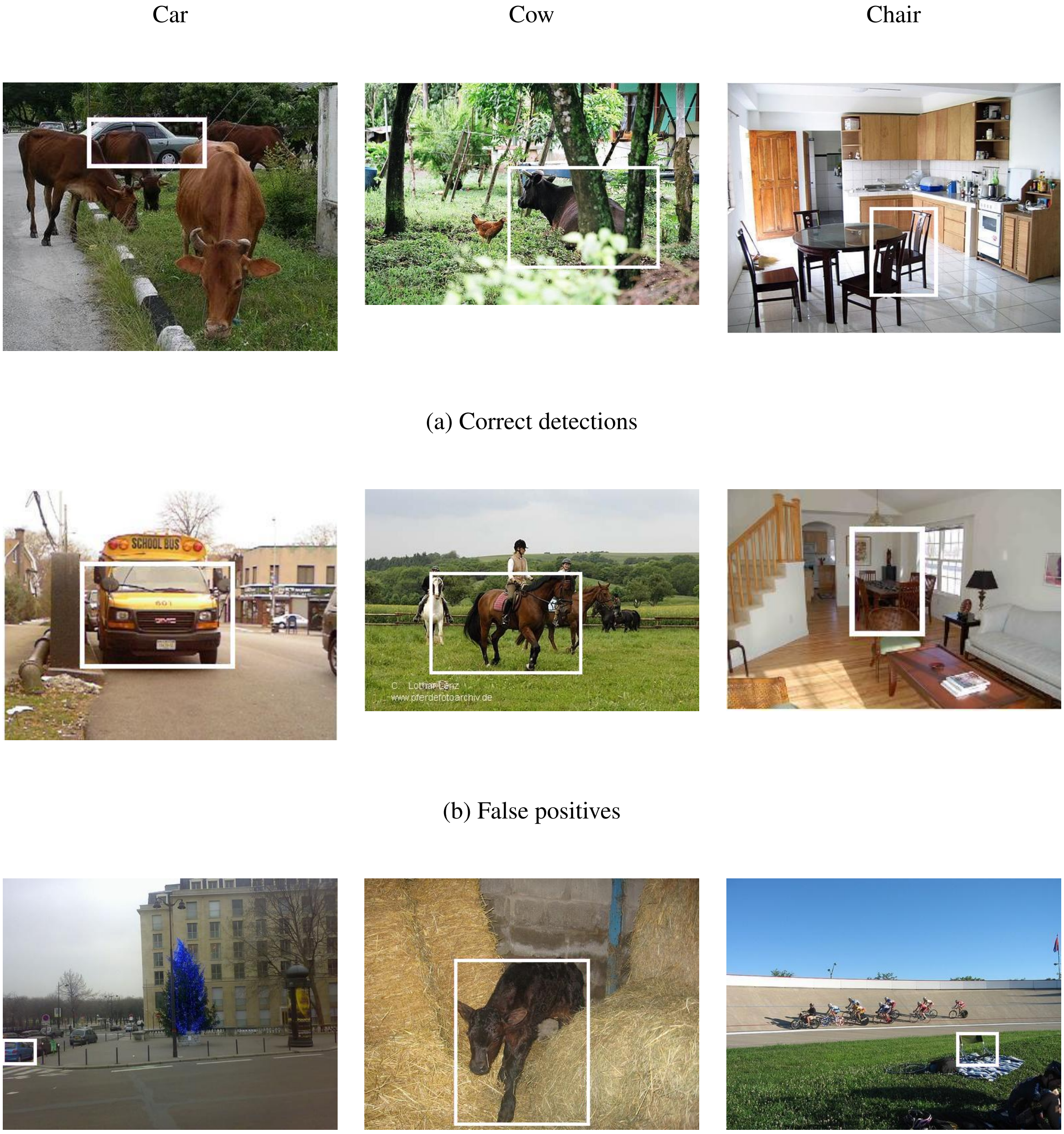
6.4.1. Learning realistic human actions from movies

Participants: Ivan Laptev [Vista, INRIA Rennes], Marcin Marszalek, Cordelia Schmid.

We address recognition of natural human actions in diverse and realistic video settings [25]. This challenging but important subject has mostly been ignored in the past due to several problems, one of which is the lack of realistic and annotated video datasets. Our first contribution is to address this limitation and to investigate the use of movie scripts for automatic annotation of human actions in videos. We evaluate alternative methods for action retrieval from scripts and show benefits of a text-based classifier. Using the retrieved action samples for visual learning, we next turn to the problem of action classification in video. We present a new method for video classification that builds upon and extends several recent ideas including local space-time features, space-time pyramids and multi-channel non-linear SVMs. The method is shown to improve state-of-the-art results on the KTH action dataset. Given the inherent problem of noisy labels in automatic annotation, we show a high tolerance of our method to annotation errors in the training set. We finally apply the method to learning and classification of challenging action classes in movies and show promising results.

6.4.2. Actions in context

Participants: Ivan Laptev [Vista, INRIA Rennes], Marcin Marszalek, Cordelia Schmid.



(a) Correct detections

(b) False positives

(c) Missed objects

Figure 9. Detection examples for the classes car, cow and chair. (a) Examples of objects correctly detected. (b) Examples of false positives with a high score. (c) Examples of objects that we were not able to detect.

We exploit the context of natural dynamic scenes for human action recognition in video. Human actions are frequently constrained by the purpose and the physical properties of the scene and demonstrate high correlation with particular scene classes. For example eating often happens in kitchens while running is more common outdoors. The contribution is three-fold: (a) we automatically discover relevant scene concepts and their correlation with human actions from the data, (b) we show how to learn selected scene concepts from video without manual supervision and (c) we develop a joint framework for action and scene recognition and demonstrate improved action recognition in natural video.

To achieve these goals we make use of movie videos and associated text scripts. For selected action classes we identify correlated scene classes using text mining. We then use script-to-video alignment to identify scene samples and to train visual scene models. Our visual models for scenes and actions are formulated within the bag-of-features framework and are combined into a joint scene-action classifier using text based co-occurrences as well as Support Vector Machines. We validate the method and report experimental results for a dataset with twelve action classes and ten scene classes acquired from 69 movies. Figure 10 shows a few examples where context helps to significantly improve action recognition.

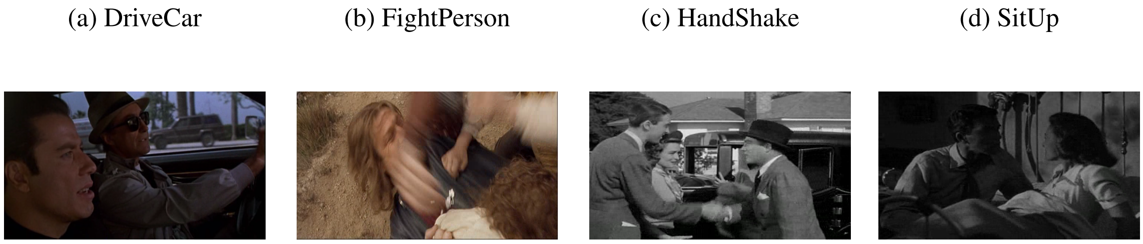

(a) DriveCar (b) FightPerson (c) HandShake (d) SitUp

Figure 10. Sample frames for action video clips where the context significantly helps recognition. Please note the car interior for driving, the outdoor setting for fighting, the house exterior for handshaking, and finally the bedroom for sitting up.

6.4.3. Spatio-temporal descriptor for action recognition

Participants: Alexander Kläser, Marcin Marszalek, Cordelia Schmid.

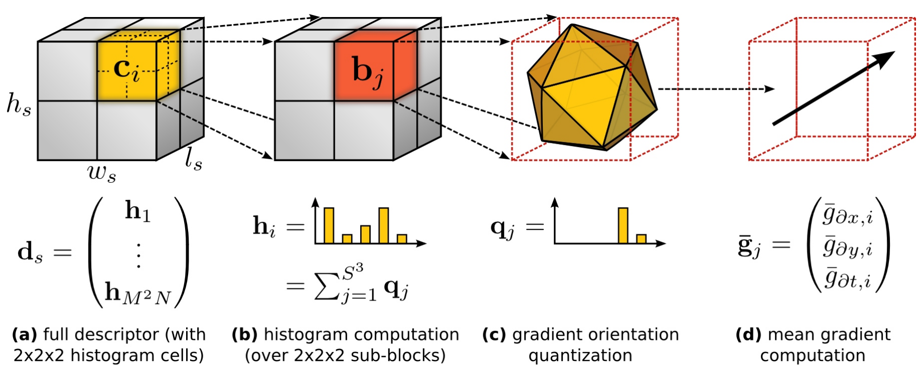

$\mathbf{c}_i$
$\mathbf{b}_j$
h_s
w_s
l_s
$\mathbf{d}_s = \begin{pmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_{M^2N} \end{pmatrix}$
$\mathbf{h}_i =$
$= \sum_{j=1}^{S^3} \mathbf{q}_j$
$\mathbf{q}_j =$
$\bar{\mathbf{g}}_j = \begin{pmatrix} \bar{g}_{\partial x,i} \\ \bar{g}_{\partial y,i} \\ \bar{g}_{\partial t,i} \end{pmatrix}$
(a) full descriptor (with 2x2x2 histogram cells) **(b)** histogram computation (over 2x2x2 sub-blocks) **(c)** gradient orientation quantization **(d)** mean gradient computation

Figure 11. Overview of the descriptor computation: (a) the support region around a point of interest is divided into a grid of gradient orientation histograms; (b) each histogram is computed over a grid of mean gradients; (c) each gradient orientation is quantized using regular polyhedrons; (d) each mean gradient is computed using integral videos.

Despite recent developments in action recognition, there exist only very few descriptors for video that incorporate spatial as well as temporal information. Figure 11 shows an overview of our novel spatio-temporal descriptor [24]. Building on the success of descriptors based on histogram of gradient orientations for static images, we view videos as spatio-temporal volumes and generalize the key concepts to the spatio-temporal domain. We propose a 3D orientation quantization based on regular polyhedrons. Furthermore, we introduce a memory-efficient strategy to compute spatio-temporal gradients for arbitrary locations and scales in constant time. The parameters of the new descriptor are evaluated on various datasets and are optimized for action recognition. A comparison with existing video descriptors is based on action classification with a bag-of-features approach that represents video sequences as orderless collections of local features types. Results show that our approach outperforms existing descriptors.

6.4.4. *Action recognition in video sequences with and without human localization*

Participants: Alexander Kläser, Marcin Marszalek, Cordelia Schmid.

Human action recognition in cluttered video scenes is a difficult task. Yet, knowledge about the localization of humans in a video sequence, provided for example by a person detector, can improve performance. Clutter can, then, be excluded and geometric constraints can be enforced. We investigate in our current research to which extent additional information, provided by the human location in a video, can improve results. In our experiments, we employ a bag-of-features based approach along with spatial grids on human tracks to exploit information provided by human tracks, see figure 12. We study systematically the influence of varying amounts of geometric constraints on the classification performance. Experiments are carried out on several challenging and realistic datasets. Results show a consistent improvement due to a prior human localization. Furthermore, our human-centered approach outperforms existing state-of-the-art methods.

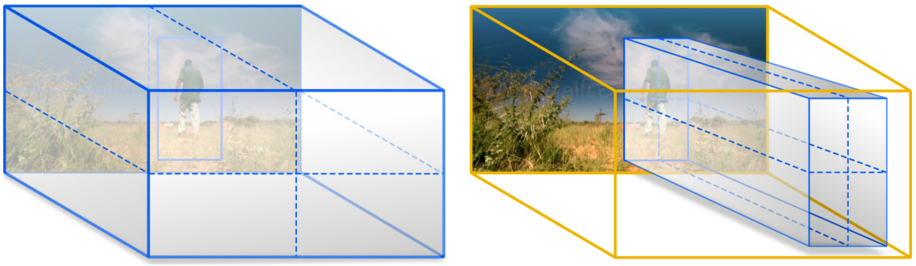


Figure 12. Action recognition is based on a bag-of-features approach; geometric information is incorporated in form of spatial grids on (left) full video sequences and (right) human tracks.

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 team in previous years. 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, see Sections 5.1 and 5.2.

Hervé Jégou and Cordelia Schmid are scientific advisors of the start-up MILPIX.

7.2. MDBD Aerospatiale

Participants: Hedi Harzallah, Frédéric Jurie, Cordelia Schmid.

There has been a collaboration with the Aerospatiale section of MBDA 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 three-year contract 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 February 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 archaeology, cultural heritage preservation, environmental science, and sociology, and it will be validated by collaborations with researchers and practitioners in these fields.

8. Other Grants and Activities

8.1. National Projects

8.1.1. QUAERO

Participants: Matthijs Douze, Hervé Jégou, Frédéric Jurie, 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, for example 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.

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 RobM@rket

Participants: Frédéric Jurie, Cordelia Schmid.

RobM@rket is an ANR “systèmes interactifs et robotique” project which started in early February 2008 for a three year period. The project aims at developing a robot system which automatically packages the items of an internet order. The robotic system is a mobile platform with an industrial arm, onto which the different algorithms, i.e., grasping, visual servoing and object detection, will be integrated. The partners are the company BA Systèmes, CEA List, the INRIA project-team Lagadic as well as the INRIA project-team LEAR and the university of Caen.

8.1.6. GRAVIT Grant

Participants: Hervé Jégou, Benoit Mordelet, Cordelia Schmid.

The GRAVIT (Grenoble Alpes Valorisation et Innovation Technologique) grant funds transfer and maturation of technology to make it usable in real-world applications. The grant started in May 2007 for a duration of 16 months. Its main goal was to extend our image indexing platform, BIGIMBAZ, to videos. This has shown very successful in the TRECVID video copy detection competition. An additional goal was to handle intellectual property issues, i.e., (1) to ensure that the components used in our platform are not patented and (2) to identify at an early stage the components of our systems that should be protected (resulting in a patent application, see section 5.1).

8.2. International Projects

8.2.1. FP6 European Project CLASS

Participants: Moray Allan, Yves Gufflet, Alexander Kläser, 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 Finland and MPI Tuebingen, Germany.

8.2.2. FP7 European Network of Excellence PASCAL 2

Participants: Adrien Gaidon, Matthieu Guillaumin, Frédéric Jurie, Marcin Marszalek, Cordelia Schmid, Jakob Verbeek.

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

8.2.3. FP6 Marie Curie EST Host Grant VISITOR

Participants: Marcin Marszalek, Cordelia Schmid.

LEAR is one of the teams participating in VISITOR, a 3 year Marie Curie Early Stage Training Host grant of the GRAVIR-IMAG laboratory. VISITOR funded the PhD of the Polish student Marcin Marszalek, from 09/2005 to 09/2008.

8.3. Bilateral relationships

8.3.1. Associated team Tethys

Participants: David Forsyth [UIUC], Martial Hebert [CMU], Akash Kushal [UIUC], Marcin Marszalek, Caroline Pantofaru [CMU], Jean Ponce [ENS Ulm], Cordelia Schmid.

The associated team Tethys started in January 2007 for two year, and has recently been extended for an additional year (2009). 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 2008, several visits of senior and junior researchers took place, see http://lear.inrialpes.fr/people/schmid/renouvellement_2008.html for details.

9. Dissemination

9.1. Leadership within the scientific community

- Conference and workshop organization:
  - C. Schmid: Organizer of 2008 International Workshop on Object Recognition.
  - F. Jurie: Workshop Chair in conjunction with ECCV'2008.
- Editorial boards:
  - C. Schmid: International Journal of Computer Vision.
  - C. Schmid: Foundations and Trends in Computer Graphics and Vision.
- Program chair:
  - C. Schmid: ECCV'2012.
- Area chairs:
  - C. Schmid: ECCV'2008.
  - C. Schmid: RFIA'2008.
  - C. Schmid: ICCV'2009.

- Program committees:
  - CVPR’2008: M. Allan, F. Jurie and J. Verbeek.
  - CVPR’2009: H. Jégou, F. Jurie, D. Larlus, M. Marszalek, C. Schmid and J. Verbeek.
  - ECCV’2008: H. Jégou, F. Jurie and J. Verbeek.
  - NIPS’2008: J. Verbeek.
  - RFIA’2008: H. Jégou and F. Jurie.
- Prizes:
  - Top results for the copy detection task of the TRECVID 2008 evaluation campaign (22 active participants), see http://www-nlpir.nist.gov/projects/trecvid/ and our notebook paper [18].
  - Winner of several of the PASCAL VOC’2008 visual object class challenges. LEAR won the detection contest for 11 out of 20 classes and the classification contest for 8 out of 20 classes, see http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008.
  - Best poster prize, honorable mention, CVPR’2008 for the paper [27].
  - Moray Allan received an Outstanding Reviewer Award at CVPR’2008.
- Other:
  - H. Jégou was a member of the “Commission de spécialistes” of the University of Grenoble, for Computer Sciences (Section 27) in 2008.
  - 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).
  - C. Schmid is a member of INRIA’s “Commission d’Évaluation”. She participated in several recruitment committees in 2008.
  - 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).

9.2. Teaching

- M. Guillaumin, Networks, pratical courses and tutorials, ENSIMAG (Bachelor-3 and Master-1), 70h.
- M. Guillaumin, Compilation and programming languages, tutorials, ENSIMAG (Master-1), 18h.
- M. Douze and H. Jégou, Multi-media databases, INPG, ENSIMAG (Master-2), 18h.
- D. Larlus, Functional programming in Scheme, tutorials, MIASS (Bachelor-3), 29h
- D. Larlus, Applied Mathematics for Computer Sciences, lecture+tutorials, Master ICA (IHS option), 11h+11h.
- C. Schmid and J. Verbeek, Machine Learning & Category Representation, Master-2, 18h.
- C. Schmid, Object recognition and computer vision, ENS Ulm, (Master-2 MVA), 10h.

9.3. Invited presentations

- M. Douze, IN’TECH Seminar, Grenoble, June 2008.
- H. Harzallah, PASCAL VOC’08 workshop, ECCV’2008, Marseille, October 2008.
- H. Jégou, Seminar “Google visiting Grenoble”, Grenoble, June 2008.

- H. Jégou, Seminar at Xerox Research Center Europe, Grenoble, July 2008.
- H. Jégou, Seminar at Ecole Normale Supérieure, Paris, September 2008.
- H. Jégou, TRECVID Workshop, Gaitherburg (USA), November 2008.
- H. Jégou, Seminar at GISPA-Lab, Grenoble, November 2008.
- H. Jégou, Seminar at Carnegie-Mellon University, Pittsburg, December 2008.
- D. Larlus, Seminar at Université de Bourgogne, Dijon, February 2008.
- D. Larlus, Seminar at the Institute des Systèmes Intelligents et de Robotique, Paris, April 2008.
- D. Larlus, Seminar at Multimodal Interactive Systems Group, TU Darmstadt, Germany, April 2008.
- D. Larlus, Seminar at the LIRIS laboratory, Lyon, September 2008.
- D. Larlus, Seminar at LISTCI, Annecy, September 2008.
- R. Mohr, Presentation at the Imaginove industrial association, October 2008.
- C. Schmid, Keynote speaker at BMVC'08, Leeds, UK, September 2008.
- C. Schmid, Seminar at Max Planck Institut Saarbrücken, Germany, September 2008.
- C. Schmid, Tutorial on images features and object recognition, Lotus Hill Summer School on Computer Vision, Ezhou, China, July 2008.
- C. Schmid, Seminar at LIAMA, Beijing, China, July 2008.
- C. Schmid, Seminar at TU München, Germany, July 2008.
- C. Schmid, Panelist on *Future Directions of Computer Vision* at IEEE CVPR 2008, http://vision.eecs.ucf.edu/PanelDiscussions.pdf.
- C. Schmid, Talk at ECCV area chair symposium, Paris, June 2008.
- C. Schmid, Talk at International Workshop on Computer Vision, Venice, Italy, May 2008.
- J. Verbeek, Seminar at Autonomous University of Barcelona, September 2008.
- J. Verbeek, Seminar at Max Planck institute for Biological Cybernetics, Tuebingen, Germany, July 2008.
- J. Verbeek, Seminar at Xerox Research Centre Europe, April 2008.

10. Bibliography

Year Publications

Doctoral Dissertations and Habilitation Theses

[1] D. LARLUS. *Création et utilisation de vocabulaires visuels pour la catégorisation d'images et la segmentation de classes d'objets*, Ph. D. Thesis, Institut National Polytechnique de Grenoble, November 2008, http://lear.inrialpes.fr/pubs/2008/Lar08.

[2] M. MARSZAŁEK. *Past the limits of bag-of-features*, Ph. D. Thesis, Institut National Polytechnique de Grenoble, September 2008, http://lear.inrialpes.fr/pubs/2008/Mar08.

[3] E. NOWAK. *Reconnaissance de catégories d'objets et d'instances d'objets à l'aide de représentations locales*, Ph. D. Thesis, Institut National Polytechnique de Grenoble, March 2008, http://lear.inrialpes.fr/pubs/2008/Now08b.

Articles in International Peer-Reviewed Journal

[4] A. Agarwal, B. Triggs. *Multilevel image coding with hyperfeatures*, in "International Journal of Computer Vision", vol. 78, n^{o} 1, June 2008, p. 15–27, http://lear.inrialpes.fr/pubs/2008/AT08.

[5] G. Blanchard, L. Zwald. *Finite dimensional projection for classification and statistical learning*, in "IEEE Transactions on Information Theory", vol. 54, n^{o} 9, September 2008, p. 4169–4182, http://lear.inrialpes.fr/pubs/2008/BZ08.

[6] P. Carbonetto, G. Dorkó, C. Schmid, H. Kück, N. De Freitas. *Learning to recognize objects with little supervision*, in "International Journal of Computer Vision", vol. 77, n^{o} 1, May 2008, p. 219–238, http://lear.inrialpes.fr/pubs/2008/CDSKD08.

[7] 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", to appear, 2008, http://lear.inrialpes.fr/pubs/2008/CLTNJ08.

[8] V. Ferrari, L. Fevrier, F. Jurie, C. Schmid. *Groups of adjacent contour segments for object detection*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", vol. 30, n^{o} 1, January 2008, p. 36–51, http://lear.inrialpes.fr/pubs/2008/FFJS08.

[9] M. Heikkila, M. Pietikainen, C. Schmid. *Description of interest regions with local binary patterns*, in "Pattern Recognition", to appear, 2008, http://lear.inrialpes.fr/pubs/2008/HPS08.

[10] H. Jegou, C. Schmid, H. Harzallah, J. Verbeek. *Accurate image search using the contextual dissimilarity measure*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", to appear, 2008, http://lear.inrialpes.fr/pubs/2008/JSHV08.

[11] D. Larlus, F. Jurie. *Latent mixture vocabularies for object categorization and segmentation*, in "Journal of Image and Vision Computing", to appear, 2008, http://lear.inrialpes.fr/pubs/2008/LJ08a.

[12] S. Malinowski, H. Jegou, C. Guillemot. *Computation of posterior marginals on aggregated state models for soft source decoding*, in "IEEE Transactions on Communications", to appear, 2008, http://lear.inrialpes.fr/pubs/2008/MJG08a.

[13] S. Malinowski, H. Jegou, C. Guillemot. *Error recovery properties and soft decoding of quasi-arithmetic codes*, in "EURASIP Journal on Applied Signal Processing", vol. 2008, 2008, http://lear.inrialpes.fr/pubs/2008/MJG08.

[14] F. Moosmann, E. Nowak, F. Jurie. *Randomized clustering forests for image classification*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", vol. 30, n^{o} 9, September 2008, p. 1632–1646, http://lear.inrialpes.fr/pubs/2008/MNJ08.

Articles in National Peer-Reviewed Journal

[15] D. Larlus, F. Jurie. *Segmentation de catégories d'objets par combinaison d'un modèle d'apparence et d'un champ de Markov*, in "Information - Interaction Intelligence", to appear 2008, http://lear.inrialpes.fr/pubs/2008/LJ08b.

International Peer-Reviewed Conference/Proceedings

[16] H. CEVIKALP, B. TRIGGS, F. JURIE, R. POLIKAR. *Margin-based discriminant dimensionality reduction for visual recognition*, in "Conference on Computer Vision and Pattern Recognition", June 2008, http://lear.inrialpes.fr/pubs/2008/CTJP08.

[17] H. CEVIKALP, J. VERBEEK, F. JURIE, A. KLÄSER. *Semi-supervised dimensionality reduction using pairwise equivalence constraints*, in "International Conference on Computer Vision Theory and Applications", January 2008, p. 489–496, http://lear.inrialpes.fr/pubs/2008/CVJK08.

[18] M. DOUZE, A. GAIDON, H. JEGOU, M. MARSZAŁEK, C. SCHMID. *INRIA-LEAR's video copy detection system*, in "TRECVID Workshop", November 2008, http://lear.inrialpes.fr/pubs/2008/DGJMS08a.

[19] D. DUCLOS, J. LONNOY, Q. GUILLERM, F. JURIE. *ROBIN: a platform for evaluating, automatic target recognition algorithms*, in "Proceedings of SPIE", vol. 6967, 2008, http://lear.inrialpes.fr/pubs/2008/DLGJ08.

[20] M. GUILLAUMIN, T. MENSINK, J. VERBEEK, C. SCHMID. *Automatic face naming with caption-based supervision*, in "Conference on Computer Vision and Pattern Recognition", June 2008, http://lear.inrialpes.fr/pubs/2008/GMVS08.

[21] H. JEGOU, L. AMSALEG, C. SCHMID, P. GROS. *Query-adaptative locality sensitive hashing*, in "International Conference on Acoustics, Speech, and Signal Processing", April 2008, http://lear.inrialpes.fr/pubs/2008/JASG08.

[22] H. JEGOU, M. DOUZE, C. SCHMID. *Hamming embedding and weak geometric consistency for large scale image search*, in "European Conference on Computer Vision", LNCS, vol. I, Springer, October 2008, p. 304–317, http://lear.inrialpes.fr/pubs/2008/JDS08.

[23] T. JIANG, C. TOMASI. *Robust shape normalization based on implicit representations*, in "International Conference on Pattern Recognition", December 2008, http://lear.inrialpes.fr/pubs/2008/JT08.

[24] A. KLÄSER, M. MARSZAŁEK, C. SCHMID. *A spatio-temporal descriptor based on 3D-gradients*, in "British Machine Vision Conference", September 2008, p. 995–1004, http://lear.inrialpes.fr/pubs/2008/KMS08.

[25] I. LAPTEV, M. MARSZAŁEK, C. SCHMID, B. ROZENFELD. *Learning realistic human actions from movies*, in "Conference on Computer Vision and Pattern Recognition", June 2008, http://lear.inrialpes.fr/pubs/2008/LMSR08.

[26] D. LARLUS, F. JURIE. *Combining appearance models and markov random fields for category level object segmentation*, in "Conference on Computer Vision and Pattern Recognition", June 2008, http://lear.inrialpes.fr/pubs/2008/LJ08.

[27] J. LIEBELT, C. SCHMID, K. SCHERTLER. *Viewpoint-independent object class detection using 3D feature maps*, in "Conference on Computer Vision and Pattern Recognition", June 2008, http://lear.inrialpes.fr/pubs/2008/LSS08.

[28] M. Marszałek, C. Schmid. *Constructing category hierarchies for visual recognition*, in "European Conference on Computer Vision", LNCS, vol. IV, Springer, October 2008, p. 479–491, http://lear.inrialpes.fr/pubs/2008/MS08.

[29] T. Mensink, J. Verbeek. *Improving people search using query expansions: how friends help to find people*, in "European Conference on Computer Vision", LNCS, vol. II, Springer, October 2008, p. 86–99, http://lear.inrialpes.fr/pubs/2008/MV08.

[30] C. Pantofaru, C. Schmid, M. Hebert. *Object recognition by integrating multiple image segmentations*, in "European Conference on Computer Vision", LNCS, vol. III, Springer, October 2008, p. 481–494, http://lear.inrialpes.fr/pubs/2008/PSH08.

[31] J. Verbeek, B. Triggs. *Scene segmentation with CRFs learned from partially labeled images*, in "Advances in Neural Information Processing Systems", vol. 20, January 2008, p. 1553–1560, http://lear.inrialpes.fr/pubs/2008/VT08.

[32] L. Yang, R. Jin, R. Sukthankar, F. Jurie. *Unifying discriminative visual codebook generation with classifier training for object category recognition*, in "Conference on Computer Vision and Pattern Recognition", June 2008, http://lear.inrialpes.fr/pubs/2008/YJSJ08.

National Peer-Reviewed Conference/Proceedings

[33] D. Larlus, E. Nowak, F. Jurie. *Segmentation de catégories d'objets par combinaison d'un modèle d'apparence et d'un champs de Markov*, in "Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle", January 2008, http://lear.inrialpes.fr/pubs/2008/LNJ08f.

Scientific Books (or Scientific Book chapters)

[34] H. Jegou, M. Douze, C. Schmid. *Recent advances in image search*, in "Emerging Trends in Visual Computing", LNCS, to appear, Springer, 2008, http://lear.inrialpes.fr/pubs/2009/JDS09.

[35] S. Lazebnik, C. Schmid, J. Ponce. *Spatial pyramid matching*, in "Object categorization: computer and human vision perspectives", to appear, Cambridge University Press, 2008, http://lear.inrialpes.fr/pubs/2008/LSP08.

Research Reports

[36] V. Ferrari, F. Jurie, C. Schmid. *From images to shape models for object detection*, Technical report, INRIA RR 6600, July 2008, http://lear.inrialpes.fr/pubs/2008/FJS08a.

[37] H. Jegou, M. Douze, C. Schmid. *Hamming embedding and weak geometry consistency for large scale image search - extended version*, Technical report, INRIA RR 6709, October 2008, http://lear.inrialpes.fr/pubs/2008/JDS08a.

[38] D. Larlus, J. Verbeek, F. Jurie. *Category level object segmentation by combining bag-of-words models and Markov Random Fields*, Technical report, INRIA RR 6668, October 2008, http://lear.inrialpes.fr/pubs/2008/LVJ08b.

References in notes

[39] H. CHUI, A. RANGARAJAN. *A new point matching algorithm for non-rigid registration*, in "Computer Vision and Image Understanding", vol. 89, n$^{\text{o}}$ 2-3, 2003, p. 114–141, http://dx.doi.org/10.1016/S1077-3142(03)00009-2.

[40] V. FERRARI, F. JURIE, C. SCHMID. *Accurate object detection with deformable shape models learnt from images*, in "Conference on Computer Vision and Pattern Recognition", June 2007, http://lear.inrialpes.fr/pubs/2007/FJS07.

[41] K. MIKOLAJCZYK, C. SCHMID. *A performance evaluation of local descriptors*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", vol. 27, n$^{\text{o}}$ 10, 2005, p. 1615–1630, http://lear.inrialpes.fr/pubs/2005/MS05.

[42] K. MIKOLAJCZYK, T. TUYTELAARS, C. SCHMID, A. ZISSERMAN, J. MATAS, F. SCHAFFALITZKY, T. KADIR, L. V. GOOL. *A comparison of affine region detectors*, in "International Journal of Computer Vision", vol. 65, n$^{\text{o}}$ 1/2, 2005, p. 43–72, http://lear.inrialpes.fr/pubs/2005/MTSZMSKG05.

[43] J. SIVIC, A. ZISSERMAN. *Video Google: a text retrieval approach to object matching in videos*, in "International Conference on Computer Vision", vol. 2, oct 2003, p. 1470–1477, http://www.robots.ox.ac.uk/~vgg/publications/html/sivic06c-abstract.html.