Active SVM-based Relevance Feedback with Hybrid Visual and representation

Marin Ferecatu, Michel Crucianu, Nozha Boujemaa

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Active SVM-based Relevance Feedback with Hybrid Visual and Conceptual Image Representation

Marin Ferecatu — Michel Crucianu — Nozha Boujemaa

N° 5558

Avril 2005

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--5558--FR+ENG

Active SVM-based Relevance Feedback with Hybrid Visual and Conceptual Image Representation

Marin Ferecatu , Michel Crucianu , Nozha Boujemaa

Thème COG — Systèmes cognitifs
Projets IMEDIA

Rapport de recherche n° 5558 — Avril 2005 — 20 pages

Abstract: Most of the available image databases have keyword annotations associated with the images, related to the image context or to the semantic interpretation of image content. Keywords and visual features provide complementary information, so using these sources of information together is an advantage in many applications.

We address here the challenge of semantic gap reduction, through an active SVM-based relevance feedback method, jointly with a hybrid visual and conceptual content representation and retrieval.

We first introduce a new feature vector, based on the keyword annotations available for the images, which makes use of conceptual information extracted from an external ontology and represented by "core concepts". We then present two improvements of the SVM-based relevance feedback mechanism: a new active learning selection criterion and the use of specific kernel functions that reduce the sensitivity of the SVM to scale.

We evaluate the use of the proposed hybrid feature vector composed of keyword representations and the low level visual features in our SVM-based relevance feedback setting. Experiments show that the use of the keyword-based feature vectors provides a significant improvement in the quality of the results.

Key-words: interactive image search, relevance feedback, SVM, active learning, conceptual descriptor, conceptual similarity

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 —Télécopie : +33 1 39 63 53 30

Retour de pertinence actif avec une représentation hybride, visuelle et conceptuelle, des images

Résumé : Dans la plupart des bases d'images professionnelles, des mots-clés sont associés aux images. Ces mots-clés concernent soit le contexte des images, soit une interprétation sémantique du contenu des images. Les mots-clés et les descripteurs visuels sont donc deux sources complémentaires d'information et il est avantageux, dans beaucoup d'applications, d'utiliser ces sources ensemble.

La question abordée dans ce rapport est la réduction du gap sémantique, grâce à une méthode active de retour de pertinence, basée sur les SVM, ainsi que par l'utilisation d'une représentation hybride, visuelle et conceptuelle, des informations fournies par une image et ses mots-clés.

Nous introduisons d'abord un nouveau descripteur, basé sur les mots-clés associés aux images, qui fait bon usage de l'information de nature conceptuelle extraite d'une ontologie externe et représentée par des "concepts noyau". Nous présentons ensuite deux améliorations du mécanisme de retour de pertinence : un critère de sélection active, avec réduction de la redondance, des exemples sur lesquels l'utilisateur doit se prononcer à chaque itération et l'emploi de fonctions noyau qui diminuent la sensibilité des SVM à l'échelle.

L'utilisation conjointe des descripteurs visuels de bas-niveau et des descripteurs conceptuels avec notre mécanisme de retour de pertinence. L'amélioration significative des résultats montre l'intérêt de l'utilisation des nouveaux descripteurs conceptuels.

Mots-clés : recherche interactive d'images, retour de pertinence, SVM, apprentissage actif, descripteur conceptuel, similarité conceptuelle

1 Problem statement

The amount of available multimedia documents has steadily increased in later years and with it the need for efficient organization and retrieval of this information when needed. Simple arrangements of items in the database and immediate lookup is no longer sufficient in a world more interested by the content than by the description tags found in most archives. These growing needs have boosted research activities in the field of content-based image retrieval (CBIR) that used to be achieved thanks to textual annotation. Hence, besides these human-based metadata (text) that usually bring semantic information, machine-based meta-data related to the physical content and its low-level features become available as information retrieval support [24], [11].

In the case of query by visual example (QBVE), the retrieval results express an overall global visual similarity, thus an approximate similarity. In this context, we may have two images with different image components ("objects") with different shapes and appearances, but remains globally similar. For some given visual queries, this leads to differences between user intention/target and the retrieved results. Starting from these observations, our community has discussed the concept of "semantic gap" through different approaches. This was also the statement of the fact that the QBVE paradigm is not able to satisfy the multiple visual search requirements [6]. There are several ways to deal with the semantic gap. One prior work is to optimize the fidelity of physical-content descriptors (image signatures) to visual content appearance of the images. The objective of this preliminary step is to bridge what we call the numerical gap. To minimize the numerical gap, we have to develop efficient images signatures (compact and visually consistent, see [27]). The weakness of visual retrieval results, due to the numerical gap, is often confusingly attributed to the semantic gap. We think that providing richer user-system interaction allows user expression on his preferences and focus on his semantic visual-content target.

Besides, if we consider either the information provided regarding the target concept or the possibilities of interaction between the user and the system, keywords and visual content appear to be complementary to each other and are valuable to rely on both of them for the retrieval of images. Simultaneously combined image and text indexing and retrieval approaches are of great interest for the semantic gap reduction and are heavily investigated even though concrete methodological advances are rare.

The extension of annotations from one visual entity (entire image, image region, etc.) to another is a first way to establish a comprehensive relation between keywords and visual content. One should note that some keywords found in manual annotations don't refer to the visual appearance, even if for some specific database they may occur for images sharing some common visual characteristics; their association with visual content can produce spurious retrieval results.

Part of the work attempting to establish a relation between keywords and visual content consists in the modeling of the visual appearance of images or of image regions corresponding to given concepts. In [12], the authors are searching for a correspondence between image *regions* and keywords that were only provided for *entire* images but refer to regions; the method is based on the development (using expectation maximization) of a joint statistical

model of the occurrence of keywords and low-level visual descriptions. Hierarchical aspect models and latent Dirichlet allocation are evaluated in [2], where the authors also study the extension of annotations to other entire images. Supervised learning is used in [1] (see also [25]) for obtaining models (Markov models or support vector machines) of the "visual content" of "atomic concepts" that can be objects, scenes or events and are associated to keywords. In [20], descriptions of image regions are directly associated to user-provided rough visual descriptions—in terms of color, position, size, shape—of concepts in an ontology.

We first mention [16], where vectorial representations are produced for the texts associated to images and *latent semantic indexing* is performed. Every image is then described both by a vector of visual features and by the latent semantic index (vector) of the text associated to the image; text-based similarity between latent semantic vectors complements the similarity defined by visual features.

The presence of joint representations (including both visual and textual features) makes *combined* search possible, often using some form of RF as in [16], [25], [29] or [30].

By marking several images as "relevant" during a relevance feedback (RF) session, the user usually defines a similarity between these images that goes beyond what can be directly obtained from low-level visual features. Considering that this similarity is related to the presence of common keywords in the annotations of some images marked as "relevant", in [29] (see also [19]) the authors link these keywords to the images top ranked by RF. A relation between the keywords and the images is thus gradually developed. In a rather analogous setting, the association of keywords to different images marked as "relevant" during an RF session serves in [30] to update similarities between these keywords; the similarity matrix can be initialized using the synonymy relations from an ontology. A "soft" extension of annotations is then performed: a keyword-based feature vector is defined for every image and contains not only keywords that directly annotate this image but also, to some degree, keywords that were found to be similar to these. The resulting similarities between keywords can capture (to some extent) general synonymy but also contextual or user-dependent synonymy and can help in dealing with homonymy. Again, keyword-based similarity complements the similarity defined by the visual features.

We are working here on databases where every image is annotated by some keywords. We introduce a new method to create a feature vector for each image using only the keywords in its annotation. We use WordNet[1] as an ontology and we derive a set of core semantic concepts linked with the keywords used for annotating the images. For each image in the database we project the keywords in its annotation on the selected core concepts obtaining a vector representation. This feature vector can be used as any other image feature vector, for enhancing the results of a query by visual example or for improving relevance feedback.

In Sec. 2.2 we introduce our new keyword-based feature vector and in Sec. 3 we present our new SVM-based relevance feedback (RF) mechanism. We put special emphasis on the selection strategy associated with the RF learner and on the choice of the kernel that produces scale invariance with respect to the distribution of the data. In Section 4 we present

[1] `http://wordnet.princeton.edu`

experimental results obtained on a real-world database from the Alinari Picture Library. We conclude this document by a summary of the main achievements of our approach.

2 Description of the images

Simultaneously combined image and text indexing and retrieval approaches are of great interest for the semantic gap reduction and are heavily investigated even though concrete methodological advances are rare. We briefly present the visual content descriptors we are using and we introduce a new keyword-based conceptual descriptor.

2.1 Visual content descriptors

For the description of the visual content of the images, we employ weighted color histograms described in [27], [5] using the Laplacian and local probability as pixel weighting functions. Weighting functions bring additional information into the histograms (e.g. local shape or texture), which is important for building compact and reliable image signatures. The resulting integrated signatures generally perform better than a combination of classical, single-aspect features.

To describe the shape content of an image we use a histogram based on the Hough transform, which gives the global behavior along straight lines in different directions. Texture feature vectors are based on the Fourier transform, obtaining a distribution of the spectral power density along the frequency axes. This signature performs well on texture images and, used in conjunction with other image signatures, can significantly improve the overall behavior. The resulting joint feature vector has more than 600 dimensions. With such a high number of dimensions, RF can become impractical even for medium-size databases and the task of the learner is also more difficult. In order to reduce the dimension of the feature vectors, we use linear principal component analysis (PCA), which is actually applied separately to each of the image features previously described.

2.2 New keyword-based conceptual descriptor

We put forward here a new, conceptual feature vector based on the set of keywords that annotate an image. This new feature vector provides complementary information both to the relevance feedback (RF) mechanism and to the evaluation of the similarity between images in a query by example (QBE) framework. With such a feature vector representation, the conceptual information brought in by the annotations can be processed by RF or QBE exactly as more classical visual feature vectors.

We previously implemented a first simple solution for representing the set of keywords associated to the images as feature vectors, that consists in using one dimension for every keyword annotating an image. Not only this solution lacks scalability, but the result of a simple distance computation between such vectors would only depend on the number of keywords shared by the two images and not on the conceptual similarities between

different keywords. Standard dimension reduction methods may provide more compact representations, but the individual dimensions in these new representations would no longer be interpretable, so the individual feature vectors would not be comprehensible any more.

In this work, to obtain a scalable solution for representing sets of keywords as comprehensible feature vectors, we suggest to select a limited set of "core" concepts and to associate to every such concept a dimension in the feature vector. We rely on an ontology, defining semantic relations between concepts, to find good candidates for these core concepts and to define the feature vectors for sets of keywords. After a brief presentation of WordNet, the general ontology we are using, we describe in this section our method for computing the conceptual feature vectors.

WordNet is a well-known and freely available general purpose ontology. English nouns, verbs, adjectives and adverbs are organized into synonym sets (synsets), each representing one underlying lexical concept. The concepts are linked by semantic relations of various types, such as synonymy, antonymy, hypernymy[2], hyponymy, holonymy, meronymy, etc. Every word has one or several meanings, corresponding to concepts in the ontology. Further details regarding WordNet can be found in [21], [3] and [13].

The **core concepts** we need for building the conceptual feature vectors should allow us to evaluate the conceptual similarity between keywords w that are mapped to different concepts $c(w)$ in the ontology. We must then rely on the hypernymy/hyponymy subgraph in WordNet linking the concepts associated to all the keywords in the database to the most generic concepts (such as "entity"). For every concept corresponding to a keyword annotating an image, we find all the paths in the ontology that lead to the most generic concepts. All the paths obtained for all the keywords in the database will define the hyponymy subgraph we are interested in. A small set (compared to the number of different keywords) of core concepts is then manually selected; good candidates are both super-concepts of several $c(w)$ concepts and are relatively close to these; also, the core concepts must be balanced among all the branches containing $c(w)$ concepts.

After selection of the core concepts, we have to compute for every image a **conceptual feature vector** representing the projection of its keywords. We first study representations for single keywords, then we turn to sets of keywords.

In all the feature vectors, one dimension is dedicated to every core concept. Suppose that $\{C_i | 1 \leq i \leq n\}$ are the n core concepts selected. Let us consider a keyword w mapped to a concept $c(w)$ and denote by $\mathbf{v}(c(w))$ the feature vector representing this keyword alone. A simple solution is to define the components of the feature vector according to

$$v_i(c(w)) = \begin{cases} 1, & \text{if } C_i \text{ is a super-concept of } c(w) \\ 0, & \text{otherwise} \end{cases}$$

This method for computing feature vectors is denoted in the following by WNS-BINARY. The keywords mapped to concepts that are different but have the same core super-concepts will have the same feature vectors. A refined solution should include in $\mathrm{v}(c(w))$ the *degrees*

[2] A concept X is a hypernym of a concept Y if Y is "a kind of" X.

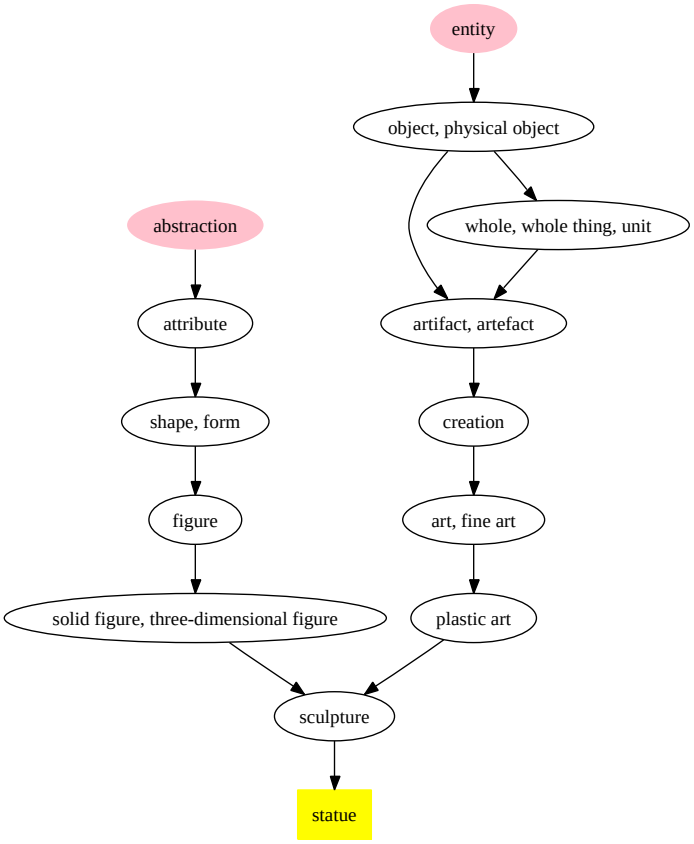

Figure 1: Hypernym graph generated by WordNet for "statue".

of similarity between $c(w)$ and its core super-concepts. Such a degree of similarity can be interpreted as the relevance of a core concept for describing an image annotated with the keyword. We thus have to evaluate the similarity between concepts.

There are several **measures of conceptual similarity**, relying on WordNet, that can be used for the definition of our keyword-based feature vector. The measures put forward in [17] and [28] rely on knowledge-rich sources (ontologies) alone, while those in [7], [18], [22] combine these sources with knowledge-poor sources (corpus statistics).

Leacock and Chodorow [17] rely on the lenght of the shorthest path following IS-A relations, $\text{len}(c_1, c_2)$, between two concepts c_1 and c_2, to measure their semantic similarity. The length of the path is scaled by the overall depth D of the concept taxonomy: $\text{sim}_{LC}(c_1, c_2) = -\log(\text{len}(c_1, c_2)/2D)$. Wu and Palmer [28] evaluate the similarity according to how close the two concepts are in the concept hierarchy, $\text{sim}(c_1, c_2) = 2N_3/(N_1 + N_2 + 2N_3)$, where

c_3 is the nearest common super-concept (or lowest super-ordinate) of c_1 and c_2, N_1 is the number of nodes in the path from c_1 to c_3, N_2 from c_2 to c_3 and N_3 from c_3 to the root node.

For Resnik [22], the similarity between two concepts depends on the extent to which they share information. The similarity between two concepts is defined as the information content of their lowest super-ordinate $\mathrm{lso}(c_1, c_2)$ according to $\mathrm{sim}_R(c_1, c_2) = -\log p(\mathrm{lso}(c_1, c_2))$, where $p(c)$ is the probability of encountering an instance of a concept c in some specific corpus. The proposal in Lin [18] is based on an information-theoretic similarity measure for arbitary objects. With the notations above, $\mathrm{sim}_L(c_1, c_2) = 2\log(p(\mathrm{lso}(c_1, c_2)))/[\log(p(c_1)) + \log(p(c_2))]$.

In a comparative study, Budanitsky and Hirst [7] present the correlations between the human rating of similarity and several similarity measures. According to their results, among the measures described above, the Lin similarity measure is closest to the way human subjects interpret semantic similarity, which is why it is our main focus in the experimental evaluations. Using one of these measures (indicated by the short names LCH, WUP, RES and respectively LIN), we defined two different types of feature vectors $\mathbf{v}(c(w))$ for representing the keyword w mapped to a concept $c(w)$. In the first one, the components of the feature vector are

$$v_i(c(w)) = \begin{cases} \mathrm{sim}(c(w), C_i), & \text{if } C_i \text{ is a super-concept of } c(w) \\ 0, & \text{otherwise} \end{cases}$$

In a second representation, we do not limit the evaluation of the similarity to the super-concepts of $c(w)$, so we set $v_i(c(w)) = \mathrm{sim}(c(w), C_i)$; in the following, the use of this method will be indicated by the "-ALL" string appended after the short name of the similarity measure employed.

If an image I is annotated with the set of keywords $\mathcal{K}(I)$, we define the components of the feature vector $\mathbf{v}(\mathcal{K}(I))$ representing $\mathcal{K}(I)$ as $v_i(\mathcal{K}(I)) = \max_{w \in \mathcal{K}(I)} v_i(c(w))$. Because of the maximum, for every core concept only the keyword that is closest to this concept has an impact on the feature vector.

In Sec. 4 we will present experiments with this new hybrid representation and show examples of retrieval together with performance evaluations.

3 SVM-based relevance feedback

Relevance feedback [31] is often used in image retrieval as a tool to refine queries or to define complex, user-dependent classes not easily described in terms of visual features. To test our approach for the joint use of visual and keyword-based image features, we use an improved scheme of SVM-based relevance feedback. First, to optimize the transfer of information between the user and the system we use a new active learning selection criterion that minimizes redundancy between the candidate images shown to the user. Second, since insensitivity to the spatial scale of the data is a desirable feature for the SVMs employed as learners, we obtain such in insensitivity by the use of specific kernel functions.

3.1 Active learning with reduction of redundancy

In order to maximize the ratio between the quality (or relevance) of the results and the amount of interaction between the user and the system, the selection of images for which the user is asked to provide feedback at the next round must be carefully studied.

Cox[10] et al. introduce some interesting ideas for the *target search* scenario, where the goal is to find a specific image in the database. The user is required to choose between the two images presented by the engine the one that is closest to the target image. The selection strategy put forward in this case attempts to identify at every round the most informative binary selections, i.e. those that are expected to maximize the transfer of information between the user and the engine (or remove a maximal amount of uncertainty regarding the target). We consider that this criterion translates into two complementary conditions for the images in the selection: each image must be ambiguous given the current estimation of the target and the redundancy between the different images has to be low. Unfortunately, the entropic criterion employed in [9], [10] does not scale well to the search of images in a larger set (*category* search) and to the selection of more than 2 images.

Tong et al. [26] present several selection criteria for SVM-learners applied to content-based text retrieval with relevance feedback. The simplest (and computationally cheapest) of these criteria consists in selecting the texts whose representations (in the feature space induced by the kernel) are closest to the hyperplane currently defined by the SVM. We call this simple criterion the selection of the "most ambiguous" (MA) candidate(s). This selection criterion is justified by the fact that knowledge of the label of such a candidate halves the version-space. While the MA criterion provides a computationally effective solution to the selection of the most ambiguous images, when used for the selection of more than one candidate image it does not remove the redundancies between the candidates.

We suggest to introduce the following additional *condition of low redundancy*: if x_i and x_j are the input space representations of two candidate images, then we require a low value for $K(x_i, x_j)$, which is the value taken by the kernel for this pair of images. If the kernel K is inducing a Hilbert structure on the feature space, if $\phi(x_i)$, $\phi(x_j)$ are the images of x_i, x_j in this feature space and if all the images of vectors in the input space have constant norm, then this additional condition corresponds to a requirement of (quasi-)orthogonality between $\phi(x_i)$ and $\phi(x_j)$ (since $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$). We shall call this criterion the selection of the "most ambiguous and orthogonal" (MAO) candidates.

The MAO criterion has a simple intuitive explanation for kernels $K(x_i, x_j)$ that decrease with an increase of the distance $d(x_i, x_j)$ (which is the case for most common kernels): it encourages the selection of unlabeled examples that are far from each other in input space, allowing to better explore the current frontier.

To implement this criterion, we first perform an MA selection of a larger set of unlabeled examples. If S is the set of images not yet included in the current MAO selection and x_i, $i = 1 \dots n$ are the already chosen candidates, then we choose as a new example the vector $x_j \in S$ that minimizes the highest of the values taken by $K(x_i, x_j)$:

$$x_j \;=\; \mathrm{argmin}_{x \in S} \max_i K(x, x_i) \qquad (1)$$

In early stages of the learning the frontier may be very unreliable and selecting those unlabeled examples that are currently considered by the learner as potentially the most relevant can sometimes produce a faster convergence of the frontier. For this reason, we added to our evaluations the following criteria: select the "most positive" unlabeled examples according to the current decision function of the SVM, denoted as MP criterion.

3.2 Kernels reducing sensitivity to scale

During the study of several groundtruth databases we found that the size of the various classes often covers an important range of different scales in the space of low level descriptors (1 to 7 in our examples). We expect yet more significant changes in scale to occur from one database to another, from one user-defined image class to another within a large database or between parts of the frontier of some classes. A too strong sensitivity of the learner to the scale of the data could then strongly limit its applicability in an RF context.

The kernels usually employed in SVM-based RF depend on a scale parameter that makes it difficult to adapt to the scale of the data. Such kernels include the RBF kernel, $K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right)$, and the Laplace kernel, $K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|\right)$. The hyperbolic kernel, $K(x_i, x_j) = 1/\left(\varepsilon + \gamma\|x_i - x_j\|\right)$, can be computed fast and we have already used it for RF with very good results.

The triangular kernel, $K(x_i, x_j) = -\|x_i - x_j\|$, was introduced in [4] as a *conditionally* positive definite kernel, but the convergence of SVMs remains guaranteed with this kernel [23]. In [15] the triangular kernel was shown to have a very interesting property: it makes the frontier found by SVMs invariant to the scale of the data. Since the triangular kernel is not positive definite but only conditionally positive definite, the account provided in [26] for the MA selection criterion does not hold for this kernel. Since the value of $K(x_i, x_j)$ decreases with an increase of the distance $d(x_i, x_j)$, our justification for the MAO criterion holds, as well as the justification of the MA criterion in [8].

In real applications, the scales of the user-defined classes cannot be known a priori and the scale parameter of a kernel cannot be adjusted online. The scale-invariance obtained by the use of the triangular kernel becomes then a highly desirable feature and experiments on several image databases prove this kernel to be a very good alternative.

4 Image retrieval relying on combined visual and semantic information

In this section we present an experimental evaluation of image retrieval using both visual features and the proposed keyword-based conceptual signature. We start by introducing the experimental setup and the performance measures we use, after which we present evaluation results both for the relevance feedback mechanism we put forward and for the joint use of visual and keyword-based feature vectors.

4.1 Experimental setup

Building the groundtruth database. We built our ground truth (GT) test database starting from an image database provided by Alinari[3]. This database has a heterogenous content, featuring images illustrating many categories of human activity, e.g. art, archeology, architecture, etc. There are 20000 images, 85601 annotations using 2059 keywords, many images being annotated by several keywords. We selected a test database having 3585 files for a total of 6664 annotations using 90 keywords. Keywords annotate between 26 and 274 images.

To have realistic classes, defined by both visual aspect and higher-level semantics, we built by hand a new ground truth. This ground truth cannot be reduced to a combination of keywords, such as no GT class is the union or intersection of sets of images annotated with the same keywords. We defined 20 classes in the GT, having between 15 and 174 images each. The number of files included in the groudtruth is 1073 and the degree of overlapping between classes is of about 10%. A certain degree of overlapping between GT classes corresponds better to real situations where an image may belong to several different user-defined image classes. While the ground truth is smaller than the database, we perform all the evaluations on the entire database of 3585 files.

The keyword-based feature vector. We built, as presented in Sec. 2.2, the hypernym graph associated with the whole test database and we choosed 28 representative core concepts to be used for projecting the sets of keywords that annotate the images. Thus, the keyword-based feature vector has 28 dimensions. No keyword was included as a core concept. To represent the visual content of the images, we use the visual feature vector presented in Sec. 2.2.

4.2 Evaluation of the relevance feedback mechanism

We evaluate the relevance feedback mechanism introduced in Sec. 3 on the test image database described above. At every feedback round the emulated user labels as "relevant" or "irrelevant" all the images in a window of size $ws = 9$. Every image in every GT class serves as the initial "relevant" example for a different RF session, while the associated initial $ws - 1$ "irrelevant" examples are randomly selected. The target of each RF session is to find all images in the GT class where the initial positive example belongs. When we use the MAO selection criterion, it is computed on a window of size $2 \times ws$.

We follow each relevance feedback session for 30 iterations (rounds)and we measure the precision within a window of size equal to the class size. This window size gives the system a chance to achieve the perfect recall, $R = 1$. Since we perform an exhaustive testing by starting a RF for each image in every class, at every iteration we compute the mean value of the precision measure over all feedback sessions. This provides a measure of how well performs relevance feedback, iteration by iteration, in its task of finding the target class. As image features, we employ a combination of the visual features and the WNS-LIN-ALL signature introduced in Sec. 2.2.

[3] `www.alinari.com`

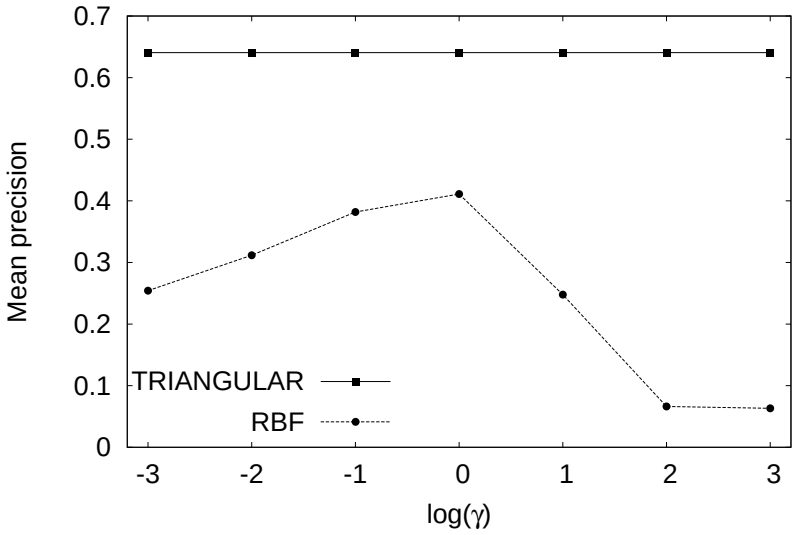

Figure 2: Mean precision vs. scale parameter for the RBF kernel and the triangular kernel.

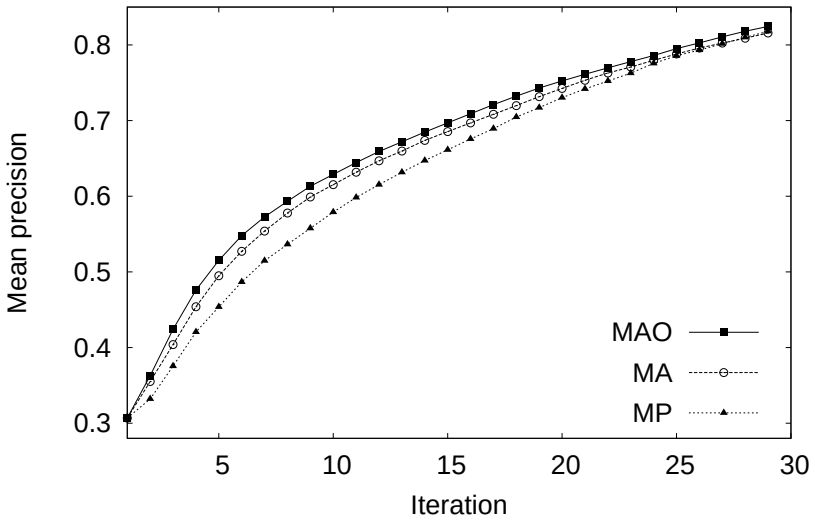

Figure 3: Comparison of several selection strategies for the triangular kernel using visual features and the WNS-LIN-ALL keyword-based signature.

First, we evaluate the sensitivity of the RBF kernel to the scale of the classes of images included in the ground truth. We use several values for the scale parameter, and for each diagram we take the mean value of the precision for the first 30 feedback iteration. This is a measure of how well performs relevance feedback with respect to the proposed GT for the given scale parameter. In Fig. 2 we present the results obtained for seven values of the scale parameter, $\gamma \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ vs. the behavior of the triangular kernel (that has no scale parameter). As we can see, the RBF kernel is very sensitive to the scale of the data. Moreover, no scale parameter value is really convenient for all classes in the ground truth, which explains why the performances of the RBF kernel are rather poor compared to those of the triangular kernel. The invariance to scale provided by the triangular kernel proves to be a very useful property for generalist databases, when the target class is complex and is best described in semantic terms.

Fig. 3 presents mean precision vs. iteration diagrams for several selection strategies: MAO, MA and MP (see Sec. 3). The MAO criterion provides better results than both MA and MP criteria. These new results both extend and confirm the evaluations presented in [14] for several GT databases and using visual descriptors alone.

4.3 Evaluation of the combined visual and conceptual descriptor

In the following we present the evaluation of the combined use of visual features and the keyword-based WNS signatures both in a QBE context and with relevance feedback. For the **QBE** situation, we test several types of conceptual feature vectors (WNS signatures) presented in Sec. 2.2 and we build precision-recall diagrams using the ground truth described previously.

In Fig. 4 we present precision/recall diagrams for the joint use of visual and WNS-LIN descriptors, the joint use of visual and WNS-LIN-ALL descriptors, and for the visual feature vector alone. The WNS-LIN-ALL signature performs clearly better than WNS-LIN when combined with the visual features, and much better than the visual feature vector alone. We obtained similar diagrams for the LCH, WUP and RES similarity measures. These findings were verified throughout the tests we performed in the QBE scenario: using both visual and keyword-based feature vectors visibly improves the quality of the results compared to using visual features alone, and projecting the keywords on all the core concepts (WNS-LIN-ALL in the figure) gives better performance than projecting only on their core super-concepts. Projecting keywords on all the core concepts allows the use of semantic relations in WordNet other than hypernymy, through the similarity functions, which has a positive influence on the results returned by the system.

We could not obtain experimental evidence to favor any of the similarity measures presented in Sec. 2.2.

We then tested the new WNS feature vectors using **relevance feedback** on our ground truth database. The comparisons were performed using the MAO selection criterion, decribed in Sec. 3, and we employed the triangular kernel for the SVM.

In Fig. 5 we compare the WNS-BINARY signatures with WNS-LIN-ALL. We see that the LIN-ALL outperforms significantly the BINARY version, both when considered alone

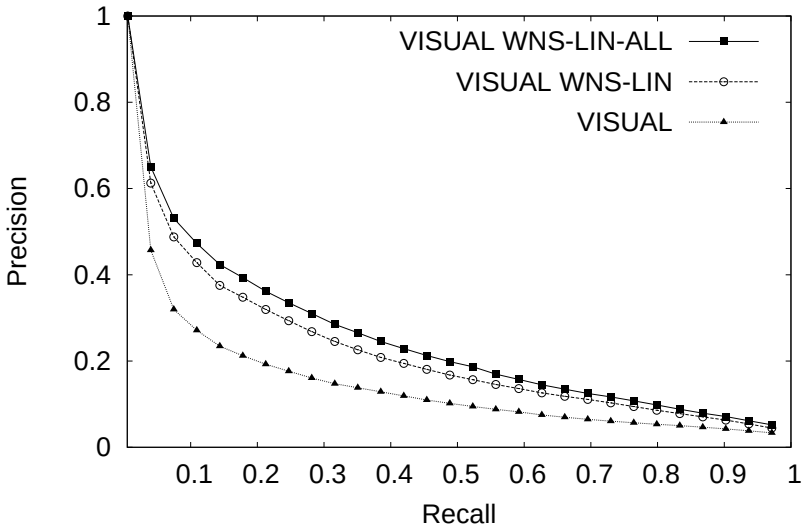

Figure 4: Precision/recall diagrams for several combinations of VISUAL, WNS-LIN and WNS-LIN-ALL descriptors.

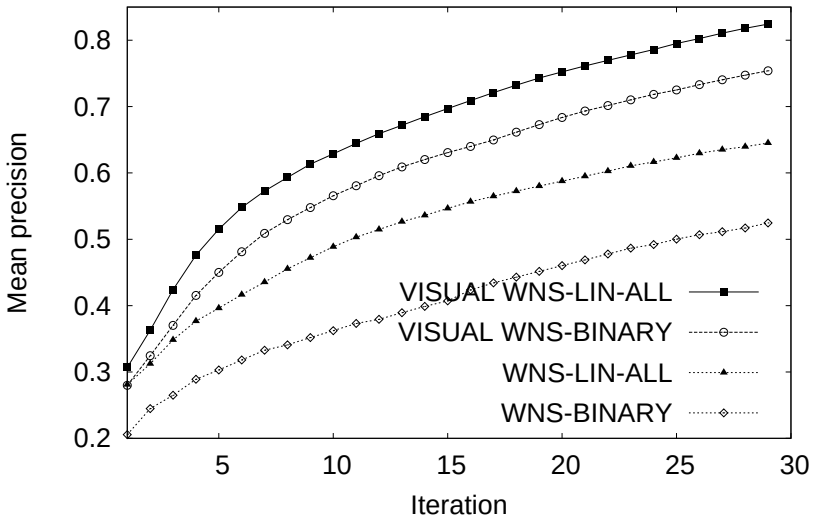

Figure 5: Comparing BINARY and WNS-LIN-ALL feature vectors with relevance feedback.

and when it is combined with the visual feature vector. Also, the joint use of visual and keyword-based feature vectors considerably improves the results compared to their individual use.

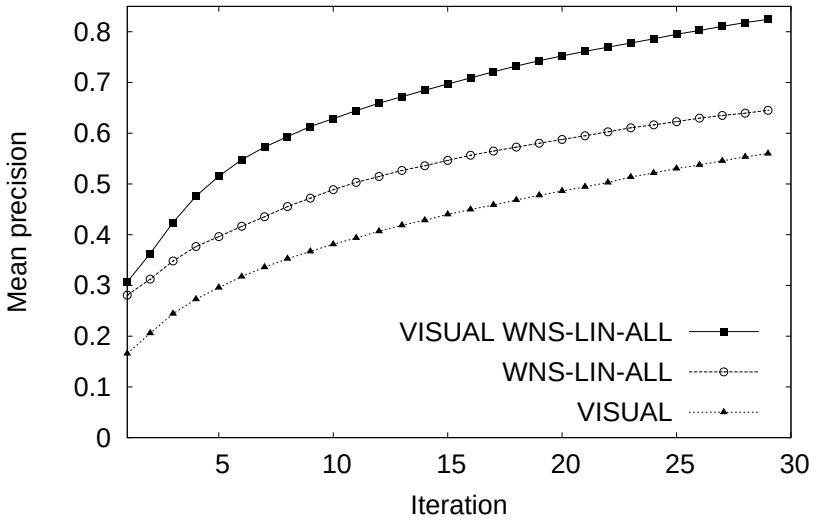

Figure 6: Mean precision vs. iteration diagrams for RF with the visual feature vector, with the WNS-LIN-ALL feature vector and with their joint use.

Fig. 6 presents mean precision vs. iteration diagrams for the WNS-LIN-ALL signature, employed alone or in combination with the visual feature vector. We see that the joint use of the keyword-based signature and visual feature vector produces a significant improvement of the results, compared to the use of visual or keyword-based signatures alone.

The tests performed with relevance feedback strongly reinforce the conclusions of the QBE evaluation: the keyword-based signatures relying on semantic similarity measures presented in Sec. 2.2 work better than the binary keyword-based signature, and projecting the keywords on all the core concepts gives better results than using only the core super-concepts. Also, the joint use of both feature vectors performs much better than using visual feature vector alone.

Moreover, the improvements obtained by using the combined feature vector were much more visible with RF than in the QBE scenario. This is an indication of the fact that user feedback allows the system to make a better use of the information provided separately by the two types of feature vectors, choosing at each iteration only what is useful in the identification of the target.

As an illustration, in Fig. 7 we present two screens of results returned by our system in a QBE scenario, the query image being in the top-left corners of the screenshots. In the left part of Fig. 7 we see the results when the system is using only the visual features; in

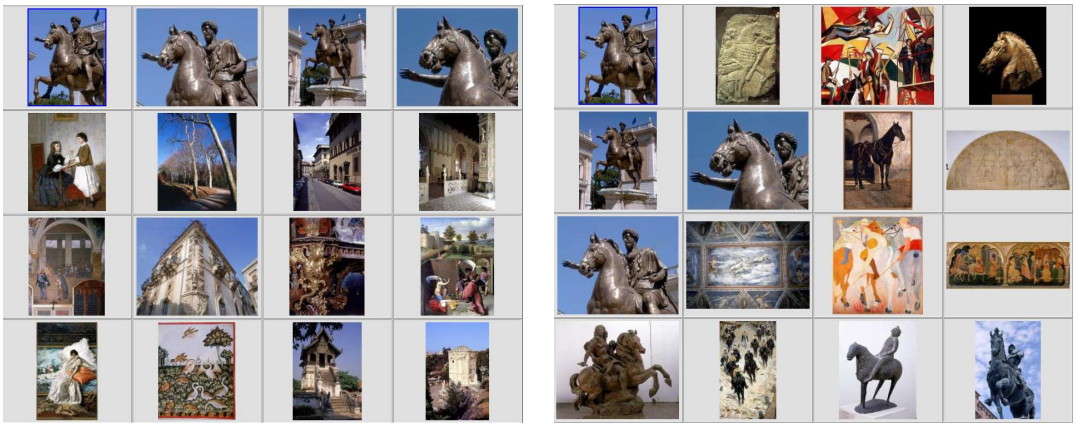

Figure 7: First page of QBE retrieval results with (left) the visual descriptor and (right) the WNS-LIN-ALL descriptor.

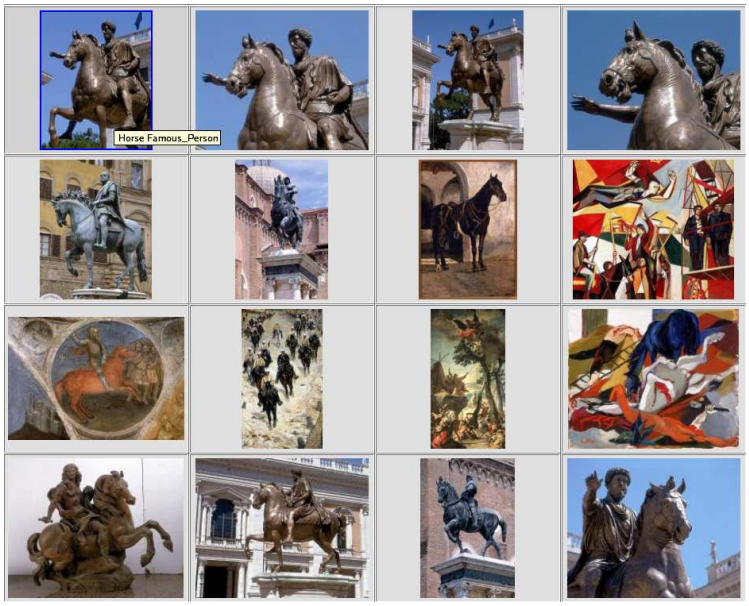

Figure 8: First page of QBE retrieval results with the combined visual and WNS-LIN-ALL descriptors.

this case the system is confused by too many images in the database having the same visual appearance as the query image. The results in the right part of Fig. 7 correspond to the use

of the WNS-LIN-ALL signature alone; while the returned images are conceptually related to the query image, their semantic content is too abstract and does not always represent well user's intent.

Fig. 8 shows the results obtained when employing both visual and keyword-based descriptors. In this case, the returned images clearly correspond better to the intent of the user.

5 Conclusion

Although image retrieval using low-level visual features works well in many situations, the problems introduced by the semantic gap limit its application to generic image databases. Alternatively, text annotations are more directly related to the high-level semantics of the images, but may not reflect visual similarities well. Keywords and visual features thus provide complementary information coming from different sources and using both of them is advantageous in many applications.

In this document we introduced a new conceptual feature vector that make use of an external ontology (WordNet) to induce a semantic generalization of the concepts corresponding to keywords. It is obviously appropriate for use with relevance feedback for large image databases. We put forward an improved relevant feedback mechanism using a new selection criterion based on active learning with reduction of redundancy between samples. We take into account the fitness of the kernel used by the SVM to the scale of the data, and we propose the use of specific kernel functions to achieve scale invariance. Evaluations performed on a ground truth build from a real generalist database confirm that our new feature vector can improve significantly the quality of the returned results, allowing an interesting reduction of the semantic gap.

6 Acknowledgements

We are very grateful to Fratelli Alinari (`http://www.alinari.com`) and especially to Andrea de Polo for providing us the annotated image database.

References

[1] W. H. Adams, Giridharan Iyengar, Ching-Yung Lin, Milind R. Naphade, Chalapathy Neti, Harriet J. Nock, and John R. Smith. Semantic indexing of multimedia content using visual, audio and text cues. *EURASIP Journal on Applied Signal Processing*, 3(2):170–185, 2003.

[2] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, March 2003.

[3] Richard Beckwith, Christiane Fellbaum, Derek Gross, and George Miller. WordNet: A lexical database organized on psycholinguistic principles. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 211–232. Erlbaum, 1991.

[4] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups.* Springer-Verlag, 1984.

[5] Nozha Boujemaa, Julien Fauqueur, Marin Ferecatu, François Fleuret, Valérie Gouet, Bertrand Le Saux, and Hichem Sahbi. Ikona: Interactive generic and specific image retrieval. In *Proceedings of the International workshop on Multimedia Content-Based Indexing and Retrieval (MMCBIR'2001)*, 2001.

[6] Nozha Boujemaa, Julien Fauqueur, and Valerie Gouet. *What's beyond query by example?* Springer Verlag, 2004.

[7] Alexander Budanitsky and Graeme Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the Workshop on WordNet and Other Lexical Resources NAACL 2001*, 2001.

[8] Colin Campbell, Nello Cristianini, and Alexander Smola. Query learning with large margin classifiers. In *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 111–118. Morgan Kaufmann, 2000.

[9] Ingemar J. Cox, Matthew L. Miller, Thomas P. Minka, Thomas Papathomas, and Peter N. Yianilos. The Bayesian image retrieval system, PicHunter: theory, implementation and psychophysical experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, January 2000.

[10] Ingemar J. Cox, Matthew L. Miller, Stephen M. Omohundro, and Peter N. Yianilos. An optimized interaction strategy for Bayesian relevance feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 553–558. IEEE Computer Society, 1998.

[11] A. del Bimbo. *Visual Information Retrieval.* Morgan Kaufmann, 1999.

[12] Pinar Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, pages 97–112. Springer-Verlag, 2002.

[13] Christiane Fellbaum and George Miller, editors. *WordNet: An Electronic Lexical Database.* The MIT Press, 1998.

[14] Marin Ferecatu, Michel Crucianu, and Nozha Boujemaa. Retrieval of difficult image classes using svm-based relevance feedback. In *Proceedings of the 6th ACM SIGMM*

International Workshop on Multimedia Information Retrieval, pages 23 – 30, October 2004.

[15] François Fleuret and Hichem Sahbi. Scale-invariance of support vector machines based on the triangular kernel. In *3rd International Workshop on Statistical and Computational Theories of Vision*, October 2003.

[16] M. La Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 24–28, 1998.

[17] Claudia Leacock, Martin Chodorow, and George A. Miller. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998.

[18] Dekang Lin. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304, 1998.

[19] Ye Lu, Chunhui Hu, Xingquan Zhu, Hong-Jiang Zhang, and Qiang Yang. A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *Proceedings of the 8th ACM International Conference on Multimedia*, pages 31–37. ACM Press, 2000.

[20] Vasileios Mezaris, Ioannis Kompatsiaris, and Michael G. Strintzis. Region-based image retrieval using an object ontology and relevance feedback. *EURASIP Journal on Applied Signal Processing*, 2004(6):886–901, June 2004.

[21] George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Five papers on WordNet. CSL Report 43, Cognitive Science Laboratory, Princeton University, July 1990.

[22] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In Chris S. Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 448–453, San Mateo, August 20–25 1995. Morgan Kaufmann.

[23] Bernhard Schölkopf. The kernel trick for distances. In *Advances in Neural Information Processing Systems*, volume 12, pages 301–307. MIT Press, 2000.

[24] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[25] John R. Smith, Sankar Basu, Ching-Yung Lin, Milind R. Naphade, and Belle Tseng. Integrating features, models and semantics for content-based retrieval. In *Proceedings of the international workshop on MultiMedia Content-Based Indexing and Retrieval (MMCBIR'01)*, pages 95–98, September 2001.

[26] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. In *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 999–1006. Morgan Kaufmann, 2000.

[27] Constantin Vertan and Nozha Boujemaa. Upgrading color distributions for image retrieval: can we do better? In *International Conference on Visual Information Systems (Visual2000)*, November 2000.

[28] Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 133–138, New Mexico State University, Las Cruces, New Mexico, 1994.

[29] Hong-Jiang Zhang and Z. Su. Improving CBIR by semantic propagation and cross-mode query expansion. In *Proceedings of the international workshop on MultiMedia Content-Based Indexing and Retrieval (MMCBIR'01)*, pages 83–86, September 2001.

[30] Xiang Sean Zhou and Thomas S. Huang. Unifying keywords and visual contents in image retrieval. *IEEE Multimedia*, 9(2):23–33, 2002.

[31] Xiang Sean Zhou and Thomas S. Huang. Relevance feedback for image retrieval: a comprehensive review. *Multimedia Systems*, 8(6):536–544, 2003.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399