Stochastic exploration and active learning for image retrieval.

Matthieu Cord, Philippe-Henri Gosselin, Sylvie Philipp-Foliguet

Stochastic exploration and active learning for image retrieval

MATTHIEU CORD, PHILIPPE H. GOSSELIN, SYLVIE PHILIPP-FOLIGUET

ETIS (CNRS UMR-8051),
University of Cergy-Pontoise,
ENSEA,
6, av. du Ponceau,
95014 Cergy-Pontoise Cedex, France
{cord,gosselin,philipp}@ensea.fr

Abstract. This paper deals with content-based image retrieval. When the user is looking for large categories, statistical classification techniques are efficient as soon as the training set is large enough. We introduce a two-step – exploration, classification – interactive strategy designed for category retrieval. The first step aims at getting a useful initial training set for the classification step. A stochastic image selection process is used instead of the usual strategy based on a similarity score ranking. This process is dedicated to explore the database in order to collect examples as various as possible of the searched category. The second step aims at providing the best classification between relevant and irrelevant images. Based on SVM, the classification applies an active learning strategy through user interaction. A quality assessment is carried out on the ANN and COREL databases in order to compare and validate our approach.

1 Introduction

Content-Based Image Retrieval (CBIR) has attracted a lot of research interest in recent years. This paper addresses the problem of category search, which aims at retrieving all images belonging to a given category from an image database.

Traditional techniques in CBIR are limited by the semantic gap, which separates the low-level information extracted from images and the semantic user request [26, 25]: the user is looking for one image or an image set with semantics, for instance a type of landscape, whereas current processing deals with color or texture features. The problem is even more complicated when the user is looking for a particular building, or a person, or for an abstract concept such as unemployment. These different levels of abstraction have been reported in [9]. Moreover, the increasing database sizes and the diversity of search types contribute to increase the semantic gap. Various strategies have been used to reduce the semantic gap.

Some off-line methods focus on the feature extraction or on the similarity function definition. Thanks to psychovisual experiments, Mojsilovic and Rogowitz [19] propose to identify image features and similarity functions which are directly connected to semantic categories. Experiments have also been carried out with user interaction to integrate a user model in a Bayesian similarity function [8]. The aim is to define a similarity between images as close as possible to the human similarity interpretation. In computer vision community, some works deal with local descriptor extraction [27] [33] and are concerned with creating indexes rotationally invariant, and robust to object deformations.

Other strategies focus on the on-line retrieval step to reduce the semantic gap. These approaches introduce human-computer interaction into CBIR [37, 36]. Interactive systems ask the user to conduct search within the database. Starting with a coarse query, the interactive process allows the user to refine the query as much as necessary. Many kinds of interaction between the user and the system have been proposed [3], but most of the times, user interaction consists of binary labels indicating whether or not the image belongs to the desired category. The system integrates these labels through relevance feedback. The main idea of relevance feedback is to use information provided by the user to improve system effectiveness.

In category search, each image has to be classified as belonging or not to the category. Retrieval techniques are mainly of two types: statistical and geometrical [36]. The geometrical methods refer to search-by-similarity systems, based on calculation of similarity between a query[1] and the images of the database [16] [24]. The objective of the statistical methods is to update a relevance function or a binary classification of images using the user labels. The approach by relevance function estimation aims at associating a score to each image of the database, expressing the

[1] Generally, one image is used as the query.

relevance of the image to the query. A Bayesian context is often used, and the probability density function is updated considering the user labels. The probability function may be uniformly initialized and iteratively refined in order to emphasize relevant images [2][8]. Recently, statistical learning approaches have been introduced in CBIR context and have been very successful [30]. Discrimination methods (from statistical learning) may significantly improve the effectiveness of visual information retrieval tasks. This approach treats the relevance feedback problem as a supervised learning problem. A binary classifier is learnt by using all relevant and irrelevant labeled images as input training data [5].

In this paper, we focus on statistical learning techniques for image category retrieval. We propose a binary classification method, but adapted to image retrieval. Indeed, the classification in CBIR context has some specificities : the input space dimension is usually very high, the training set is very small in comparison with the test set (the whole database), unlabeled data are available, *etc*. To take into account these properties, our strategy is based on Support Vector Machines (SVM) classification and on an active learning strategy [6].

In addition, learning algorithms need enough initial training data in order to get correct classification [30]. We introduce in this paper a first step based on a database exploration scheme to initialize the classification of the second step. A discrete probability law is proposed to express the relevance of images. A sampling is then applied to display new images to the user.

The main originality of our work can be summarized in two points:

-Stochastic exploration strategy to get useful initial training set;

-Active learning scheme to boost the classification step.

After an overview of the learning architecture (section 2), the main components of our strategy will be detailed (sections 3, 4 and 5). Experiments are provided in section 6 to validate and compare the strategy with up-to-date concurrent methods.

2 Retrieval system architecture

As explained in introduction, our image retrieval strategy, called RETIN, is organized in a two-stage sequential process: – 1) database exploration – 2) active learning classification. The RETIN algorithm starts a category search with the exploration strategy step before automatically switching to the classification step. The whole scheme of our retrieval strategy is reported on fig. 1.

For the exploration step (detailed in section 4), a discrete probability law expresses the relevance of images. A stochastic sampling of this law is applied to present new im-

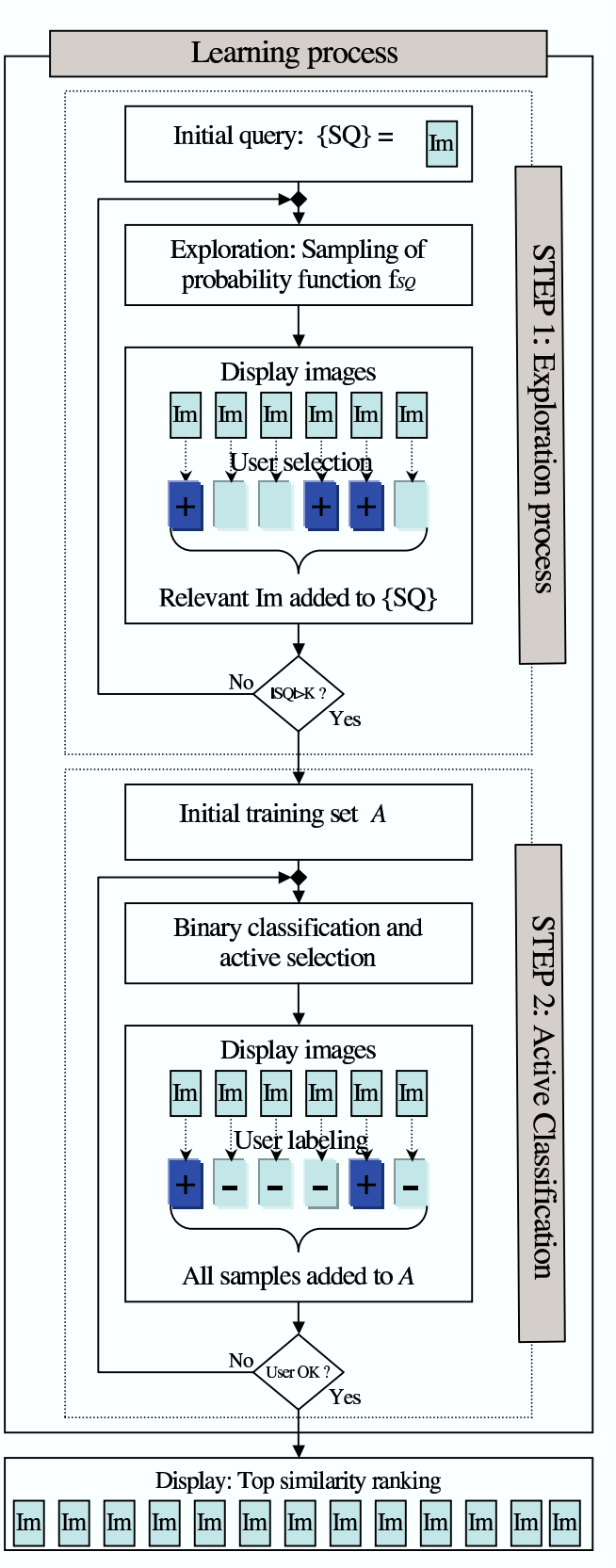

Figure 1: RETIN architecture

ages to the user. After user's labeling, new relevant images are added to the initial query and the process is iterated. The set of relevant images accumulated during this stage is called the *semantic query* ($\mathcal{SQ}$).

The exploration process stops when the cardinal $|\mathcal{SQ}|$ of the semantic query is higher than a threshold K, and the active classification process starts. $K = 20$ has been used for all our experiments to keep the same values than the ones used by other methods with which we make comparisons.

For the active classification process (detailed in section 5), we use a SVM binary classifier with specific kernel function, and a specific active learning process to sample new images to display for labeling. After user's labeling, the user decides whether to stop or to continue the learning process. If it continues, the new examples are added to the training set and the classification process is iterated. Else, the learning process is over, and the final top similarity ranking is presented to the user.

An example of the RETIN interface is reported on Fig. 2. The lower window displays the images to label during the learning scheme. The upper one is the final window, where images are displayed using top relevance ranking.

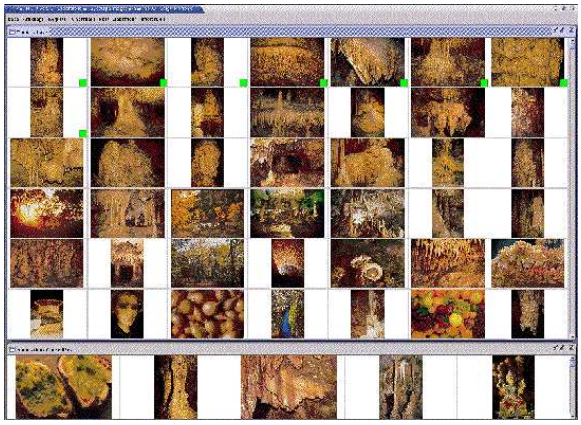

Figure 2: Example of RETIN results

Both steps need a similarity function to compare images. We introduce in section 3 statistical tools to deal with similarity and classification functions.

3 Statistical tools for CBIR

Similarity functions are usually employed to compare two images. When dealing with a more complex query (for example, a set of images), the similarity concept has to be extended. In a previous work [10], we proposed a merging scheme to combine two-by-two measures between the current image and all the relevant labeled images. Another way to estimate this similarity is to consider the problem as a probability density function estimation problem. By this way, it is easier to deal with multi-modal distributions.

Besides, when dealing with a query set having relevant and irrelevant images, a decision function used for discrimination has to be computed. Statistical learning techniques such as nearest neighbors [15], Support Vector Machines [31, 5], bayes classifiers [36], have been used.

SVM has demonstrated capacity in pattern recognition, and more recently in CBIR [31, 38]. We have shown that the SVM classification method is highly adapted to the image retrieval context [13]. This classification method can deal with high dimensionality using the *kernel trick*, and does not require a too large training set. Thus, we use SVM as our default classification method for CBIR.

In the following section, we outline the SVM algorithm, which provides a decision function [34]. For the density estimation, we also present an adaptation of SVM two-class formalism to one-class formalism [28]. This one-class SVM will be used in our exploration process.

3.1 Support Vector Machines

The Support Vector Machines (SVM) are a type of learning algorithms developed in the 1990s. They are based on results of statistical learning theory introduced by Vapnik [34]. These learning machines use kernels, which are a central concept for a number of learning tasks.

First, we assume that both classes are linearly separable. Let $(\mathbf{x}_i)_{i\in\{1,...,N\}}$, $\mathbf{x}_i \in \mathbb{R}^p$ be the feature vectors representing the training data, and $(y_i)_{i\in\{1,...,N\}}$, $y_i \in \{-1, 1\}$ be their respective class labels. Let us define a hyperplane by $< \mathbf{w}, \mathbf{x} > +b = 0$ where $\mathbf{w} \in \mathbb{R}^p$ and $b \in \mathbb{R}$. Since the classes are linearly separable, we can find a function f, $f(\mathbf{x}) =< \mathbf{w}, \mathbf{x} > +b$ with:

$$y_i f(\mathbf{x}_i) = y_i(< \mathbf{w}, \mathbf{x}_i > +b) > 0,\ \forall i \in \{1, ..., N\} \quad (1)$$

The decision function may be expressed as $f_d(\mathbf{x}) = sign(< \mathbf{w}, \mathbf{x} > +b)$ with $f_d(\mathbf{x}_i) = sign(y_i),\ \forall i$.

Since many functions realize the correct separation between training data, additional constraints are used. SVM classification method aims at finding the *optimal* hyperplane based on the maximization of the *margin*[2] between the training data for both classes.

Because the distance between point $\mathbf{x}$ and the hyperplane is $\frac{yf(\mathbf{x})}{||\mathbf{w}||}$, the optimization problem may be expressed as the following minimization:

$$\min \frac{1}{2}||\mathbf{w}||^2 \ \text{ subject to } \ y_i(< \mathbf{w}, \mathbf{x}_i > +b) \geq 1,\ \forall i \quad (2)$$

The *support vectors* are the training points for which we have an equality in Eq. 2. All of them are equally close

[2]The margin is defined as the distance from the hyperplane of the closest points, on either side.

to the optimal hyperplane. One can prove that they are sufficient to compute the separating hyperplane.

This is a convex optimization problem (quadratic criterion, linear inequality constraints). Usually, the dual formulation is favored for its easy solving with standard techniques. With $\boldsymbol{\alpha}^\star$ the dual solution of the quadratic optimization, the hyperplane decision function can be written as:

$$f_d(\mathbf{x}) = sign(\sum_{i=1}^{N} y_i \alpha_i^\star < \mathbf{x}, \mathbf{x}_i > + b)$$

The linear SVM classifier previously described finds linear boundaries in the input feature space. To get much more general decision surfaces, the feature space may be mapped into a larger space before achieving linear classification. Linear boundaries in the enlarged space are equivalent to nonlinear boundaries in the original space. Everything about the linear case also applies to nonlinear cases, using a suitable kernel k instead of the Euclidean dot product. The decision function is:

$$f_d(\mathbf{x}) = sign(\sum_{i=1}^{N} y_i \alpha_i^\star k(\mathbf{x}, \mathbf{x}_i) + b)$$

To get a relevance function useful in CBIR, the distance to the boundary is used:

$$f(\mathbf{x}) = \sum_{i=1}^{n} y_i \alpha_i^\star k(\mathbf{x}, \mathbf{x}_i) + b \tag{3}$$

Various kernel functions $k(,)$ have been proposed. The most popular ones are the Gaussian and polynomial kernels. Because we have no prior assumption on input data configuration, we selected a Gaussian kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = exp(-\frac{d^2(\mathbf{x}_i, \mathbf{x}_j)}{2\sigma^2}) \tag{4}$$

3.2 One-class SVM

When only relevant images are considered, a classification can not be carried on, but a density function may be estimated. The one-class SVM method [28] estimates the density support of a vector set $X = (\mathbf{x}_i, y_i = 1)_{i \in \{1,\dots,N\}}$ representing an image class. As SVM, this leads to a quadratic optimisation problem, and can be used with a kernel function. The same function f (Eq. 3) is computed in the one-class context, providing a density function.

4 Exploration process

Statistical learning approaches perform binary classification. They need enough training data to give acceptable results.

In our learning scheme, we propose an exploration process in order to get an initial training set for the classification process. These images have to be carefully selected because when the user is looking for large and complex categories, relevant images are usually scattered in the feature space. The system needs an exploration strategy able to efficiently catch complex categories with multi-modal distributions.

In Bayesian framework as proposed by [8] or [12], some kind of exploration is implicitly performed, but the goal is not to retrieve large categories, and exploration is not explicitly settled. We introduce in this article a stochastic scheme to manage database exploration. A probability function is defined to express the relevance of each image. The tuning of the probability law is the key point of such a strategy. It has to change from a tolerating law (large exploration) to a selective law. The proposed scheme is working with discrete probabilities to handle the law parameters.

In the next section, the image sampling principle is explained, and the parameter tuning is presented in the following one.

4.1 Sampling process

Thanks to the relevance function f (we use the probability density function estimated with the one-class SVM method), all the images may be sorted. In interactive retrieval, the basic approach selects and presents to the user the m first images according to relevance ranking. This strategy is efficient to refine similarity around positives examples of $\mathcal{SQ}$, but there is no consideration about the confidence that we can have on f during the process.

The introduction of a random scheme may be useful to explore the database and to dynamically express the confidence that we have in the semantic query model $\mathcal{SQ}$ and consequently in f (computed from $\mathcal{SQ}$).

We introduce an exploration process by using each $f_k = f(\mathbf{x}_k)$ as a weight, and find new images $\mathbf{x}_i$ such as f_i is sampled from the multinomial law $\mathcal{M}()$:

$$f_i \sim \mathcal{M}(f_1, \cdots, f_n)$$

By this way, any image from the database may be selected, even an image far from the current $\mathcal{SQ}$.

Actually, the strategy for computing the weights associated to each image may be changed in order to take into account the confidence in $\mathcal{SQ}$. It can be done using the following probability on weights f_k:

$$p_k = P(f_k) = \frac{1}{Z_T} \times \exp(\frac{f_k}{T})$$

where Z_T is the sum of the exponential values over all the images of the database and T, the parameter which tunes the confidence that we have in $\mathcal{SQ}$.

At each iteration of the interactive search, the system samples and displays images according to these probabilities : $\mathcal{M}(p_1, \cdots, p_n)$. All the images that the user labels as relevant are added to set $\mathcal{SQ}$.

Let us explain the idea of this stochastic strategy: when parameter T is high, the influence of f is weak, and thus, all weights p_k are almost equal. The sampling is then a pure random sampling. The confidence in f (and $\mathcal{SQ}$) is very low and the exploration is favored. When parameter T decreases, the influence of f increases in the probability computation. The search space cuts down around $\mathcal{SQ}$, and the confidence in $\mathcal{SQ}$ increases.

The crucial point is the tuning of T. In a previous work [7], we proposed to handle the exploration by using relation inspired by simulated annealing techniques and calculation by approximations in the continuous domain. We propose here a new formalism established on discrete probabilities. Simple assumptions allow us to make the tuning fully automatic.

4.2 Exploration tuning

First, we propose to measure the confidence c in $\mathcal{SQ}$ by using the number of images in the set: $|\mathcal{SQ}|$. Indeed, as $\mathcal{SQ}$ is composed of relevant images, the confidence c increases as the cardinal number $|\mathcal{SQ}|$. When $|\mathcal{SQ}|$ is small, the semantic query is poor and the confidence is low. When $|\mathcal{SQ}|$ is large, the set contains many images and should be rich enough to stop the exploration; the confidence in $\mathcal{SQ}$ is high. c may be expressed as $c = g(|\mathcal{SQ}|)$ where $g()$ is an increasing monotonous function. We adopt the basic relation

$$c = |\mathcal{SQ}| \tag{5}$$

Besides, the confidence c has to be linked to probability P. Let us consider the mathematical expectation of the f_k values, estimated by: $f_{expe} = \frac{1}{N}\sum_{k=1}^{N} f_k$ and f_{max} the greatest value of f_k on the database. Using exponential laws, it is possible to tune the decreasing of the probability law thanks to both $P(f_{max})$ and $P(f_{expe})$ values. Our assumption is that the ratio between these probabilities may be linked to c. Indeed, when c is low, $\frac{P(f_{max})}{P(f_{expe})}$ should be low too (around 1) to explore a lot, and when c is high, $\frac{P(f_{max})}{P(f_{expe})}$ should be very large to tighten up around the relevant images of $\mathcal{SQ}$. The relation is then as follows:

$$c = \frac{P(f_{max})}{P(f_{expe})} = \frac{\exp(\frac{f_{max}}{T})}{\exp(\frac{f_{expe}}{T})} \tag{6}$$

From equations 5 and 6 (for $|\mathcal{SQ}| > 1$), it follows[3]:

$$T = \frac{f_{max} - f_{expe}}{ln(c)} = \frac{f_{max} - f_{expe}}{ln(|\mathcal{SQ}|)}$$

[3]This condition is always true except at the beginning, where a particular condition ($c = 2$) may be used.

One step of the exploration process may be summarized as follows:

1. User's labeling; add relevant images to $\mathcal{SQ}$
2. Compute $|\mathcal{SQ}|$ and $f_k = f(x_k) \ \ \forall x_k$
3. Compute $f_{expe} = \frac{1}{N}\sum_{k=1}^{N} f_k$ and $f_{max} = max_k\{f_k\}$
4. Compute $T = \frac{f_{max} - f_{expe}}{ln(|\mathcal{SQ}|)}$
5. Compute $\forall x_k$ (unlabeled) $p_k = \frac{1}{Z_T} \times \exp(\frac{f_k}{T})$
6. Sample and display new images with $\mathcal{M}(p_1, p_2, \cdots)$

This approach allows us to have a straight control of the exploration with simple and intuitive assumptions to automatically tune the parameters.

5 Active classification process

The classification process is the second step of our two-step sequential process for category retrieval (cf. fig. 1). All the images in $\mathcal{SQ}$ are used with the irrelevant labeled images[4] in a classification framework.

The CBIR context defines a very specific classification problem. In this article, we are dealing with the following characteristics:

1. *High dimensionality*. Database images are usually represented by vectors of high dimensionality.
2. *Few training data*. As the user cannot be asked for labeling thousands of images, the system has to sort out a very small percentage of labeled data.
3. *Interactive learning*. Usually, in classification framework the training set is fixed. In interactive retrieval context, the training set grows step by step. All the images labeled during the current interaction are added to the training set for the next classification step. In statistical learning, this property defines the active learning framework [6].

After an introduction to active learning strategies and CBIR learning context, we explain our classification strategy exploiting the above mentioned specificities.

5.1 Statistical learning strategies for CBIR

Due to the first characteristic, that is to say vectors of high dimensionality (for instance, 50 or more), artifacts appear, known as the curse of dimensionality [15]. However, with

[4]All examples that the user had labeled as irrelevant during the exploration were also stored to be exploited during this second step.

the theory of kernel functions, one can reduce this problem [29], especially if kernel functions can be adapted to a specific application. For instance, when distributions are used as feature vectors, a gaussian kernel gives excellent results in comparison to distance-based techniques [13]. We use this kernel associated to SVM (see section 3) to compare images and compute classification.

Concerning the second characteristic, although there are few training data, all the unlabeled images of the database are available. Semi-supervised techniques use labeled and unlabeled images to compute the classification function, as for instance, the Transductive SVM method [17], the semi-supervised Gaussian mixtures [20], and semi-supervised Gaussian fields [39]. However, TSVM and SSGM do not lead to significant improvements [4, 14]. Furthermore, these techniques have high computational needs in comparison to inductive techniques. For now, semi-supervised learning techniques do not seem to be adapted to the context we are focusing on.

Active learning is another solution to deal with the lack of training data. The principle is that the training data set is no more fixed but new samples are phased in thanks to user interaction. Active learning strategies aim at selecting samples that, once added to the training set, will allow to optimize the classification, as for instance by minimizing the expected error of the learner [23].

Of course, the expected error is not accessible and approximation schemes have been proposed [23]. The idea is to restrict the computing of the error to the set of unlabeled data available, and to train as many classifiers as there are unlabeled data and labels. As the label of each candidate is unknown. Roy and McCallum compute the expectation for each possible label. This kind of approach is very time consuming and has never been successfully used in CBIR.

Uncertainly-based sampling is another way to perform active learning principle. The method selects the documents for which the classification function is the most uncertain. A first solution consists in selecting the unlabeled documents with the probabilities closest to 0.5 [18]. Similar strategies have been also proposed with SVM classifier [21], with a theoretical justification [31]. This strategy rests on a strong assumption: a reliable estimation of the boundary between classes. In classification framework, the training data set approximatively represents 50% of the whole data set. In CBIR, the training set stays very small (even after interaction) in comparison to the database size. In such a context, to get a reliable estimation of the boundary is a major problem. In this particular context, statistical techniques are not always the best ones, and we propose in the next section an heuristic-based correction to the estimation of f close to the boundary.

5.2 Active RETIN method for image set selection

Let $(\mathbf{x}_i)_{i\in\{1,...,n\}}$, $\mathbf{x}_i \in \mathbb{R}^p$ be the feature vectors representing images from the whole database, and $\mathbf{x}_{(i)}$ the permuted vectors after a sort according to the function f (Eq. 3). At feedback iteration j, SVM_{active} proposes to label m images from rank s_j to s_{j+m-1}:

$$\underbrace{\mathbf{x}_{(1),j}}_{\text{most relevant}}, \mathbf{x}_{(2),j}, ..., \underbrace{\mathbf{x}_{(s_j),j}, ..., \mathbf{x}_{(s_{j+m-1}),j}}_{\text{images to label}}, ..., \underbrace{\mathbf{x}_{(n),j}}_{\text{less relevant}}$$

In SVM_{active} strategy, s_j is selected so that $\mathbf{x}_{(s_j),j}, ..., \mathbf{x}_{(s_{j+m-1}),j}$ are the closest images to the SVM boundary. The closer to the margin an image is, the less its classification is reliable.

We introduce a method based on the same principle than SVM_{active}, but without using the SVM boundary to find the value s. Indeed, we notice that, although the boundary changes a lot during the first iterations, the ranking operation is quite stable. Actually, we just suppose that the best s (corresponding to the searched boundary) allows to present as many relevant images as irrelevant ones. Thus, if and only if the set of the selected images is well balanced (between relevant and irrelevant images), then s_j is good. We exploit this property to adapt s during the feedback steps.

At the jth feedback step, the user gives new labels for images $\mathbf{x}_{(s_j),j}, ...\mathbf{x}_{(s_{j+m-1}),j}$. Let us note $r_{rel}(j)$ and $r_{irr}(j)$ the numbers of relevant and irrelevant labels. To obtain balanced training sets, s has to be increased if $r_{rel}(j) > r_{irr}(j)$, and decreased otherwise. We adopt the following upgrade rule for s_{j+1}: $s_{j+1} = s_j + k \times (r_{rel}(j) - r_{irr}(j))$ For now, we have used this relation with $k = 2$ in all our experiments.

Once s_{j+1} is computed, the system should propose to the user the m images from $\mathbf{x}_{(s_{j+1}),j+1}$ to $\mathbf{x}_{(s_{j+1}+m-1),j+1}$. Actually, we also want to increase the sparseness of the training data. Indeed, nothing prevents an image close to another (already labeled or selected) to be selected. To overcome this problem, we consider exactly the same strategy but working no more on images but on clusters of images: we compute m clusters of images from $\mathbf{x}_{(s_j),j}$ to $\mathbf{x}_{(s_{j+M-1}),j}$ (where $M = 10 \times m$ for instance), using an enhanced version of LBG algorithm [22]. Next, the system selects for labeling the most relevant image in each cluster. Thus, images close to each other in the feature space will not be selected together for labeling.

6 Experiments

6.1 RETIN features and parameters

RETIN is a new version of the CBIR system developed in ETIS laboratory [11].

Color and texture information are exploited. As none of the numerous color spaces has proved its superiority

over the others for image coding, we have chosen the HSV space. For texture analysis, Gabor filters are used, with twelve different scales and orientations.

Signatures are statistical distributions of colors and textures resulting from a dynamic quantization of the feature spaces. That means we use color and texture space clustering to compute the image histograms. Both spaces are clustered using an enhanced version of LBG algorithm [22]. The main problem is to choose the number of clusters, which gives the number of bins in the histograms.

Some theoretical rules may be used to set the number of histogram bins. Sturges's or Scott's rules cited in [1] allow to avoid over or under-quantization. In image retrieval context, Brunelli and Mich have evaluated many feature histograms and they concluded that low-resolution histograms (with small bin numbers) are reliable [1]. For color histograms, Tran and Lenz suggest to use around 30 bins [32]. In a previous paper [11], we made a lot of comparisons using different numbers of clusters for dynamic and static quantizations of the feature space, which all confirm these propositions. A major advantage of the dynamic approach is the reduction of the size of the signature without performance degradation. For a generalist database (around $10,000$ images), a small number of classes obtained by a dynamic clustering of the database is sufficient to build efficient signatures. We have adopted this dynamic quantization in the RETIN system with 25 classes (as the default value).

Image signature consists of one vector representing the image color and texture distributions. The input size p is then 50 in our experiments.

The kernel function used in the SVM algorithm (one-class and two-class) is a Gaussian kernel (Section 3). Moreover, the distance in Gaussian kernel may be chosen according to feature vector type. We use a χ^2 distance which is well suited for vectors representing distributions, and in that case, $\sigma = 1$ (Eq. 4).

6.2 Evaluation and comparison protocol

6.2.1 Databases

The tests are carried out on two generalist databases: the *ANN* [5] image database and the COREL photo database. *ANN* contains around 500 images divided into 11 categories from 25 to 50 images.

The COREL database contains more than $50,000$ pictures organized in categories. Each category has about 100 images. To get tractable computation for the statistical evaluation, we randomly selected 10% of the COREL categories. We obtained about 50 categories and the corresponding database is composed of $6,000$ images. We present here results for 5 categories directly extracted from the initial 50 categories or obtained by merging some of them (to get sets with different sizes and complexities). The important point is to show results from small and mono-modal categories to large and multi-modal categories. They are reported in table 1 from the simplest one to the most complicated one.

category	size	description
caverns	121	simple, mono modal
doors	199	simple, rather mono modal
flowers	506	very large, few modes
savanna	399	large, few modes
landscape	451	complicated, many modes

Table 1: COREL categories for evaluation

6.2.2 Statistical performance measurements

The CBIR system performances are measured using precision(P), recall(R) and statistics computed on P and R for each category. Let us note A the set of images belonging to the category, and B the set of images returned to the user, then: P $= \frac{|A \cap B|}{|B|}$ and R $= \frac{|A \cap B|}{|A|}$. Usually, the cardinality of B varies from 1 to database size, providing many points (P,R). We present P/R curves which may be displayed using interpolation of the (P,R) points.

To carry out quantitative evaluation, we use the breakeven point bp metric. The breakeven point is defined as the point on a precision-recall curve that has the same value for precision and recall. There is an obvious relation between a breakeven point and the performance of a classification or retrieval system: $|A| = |B|$. It is a very interesting measure for comparison purposes when looking for large categories. We also use the Mean Average Precision (MAP) which represents the value of the P/R integral function. This metric is used in the TREC VIDEO conference[6], and gives a global evaluation of the system (over all the (P,R) values).

Each simulation is initialized with one image randomly selected within the desired category. For each feedback step, m images are automatically labeled using the ground truth. The training stops after $i = 10$ iterations in these experiments. 100 simulations are done for each category, and P and R average values are computed.

6.2.3 Comparative methods

For the quality assessment, our strategy RETIN has been compared with three methods:

[5] "Labeled ground-truth database", *Department of Computer Science and Engineering, University of Washington*, http://www.cs.washington.edu/research/imagedatabase/.

[6] http://www-nlpir.nist.gov/projects/trecvid/

M1: a SVM classification algorithm without exploration or active strategy. It means that we use a learning data set where images are randomly selected in the database. Of course, the same number of training data is used.

M2: a reference classification-based strategy for relevance feedback. We use a Bayesian classifier with a Parzen window density estimation according to the framework of Vasconcelos [35].

M3: a reference active strategy learning, the SVM_{active} strategy [31].

All the methods follow the same interactive protocol and do not require any manual tuning before the process.

6.3 Results

6.3.1 ANN

Because of the small size of the database, the number m of images labeled at each interactive feedback step is set to $m = 5$. The number of feedbacks is set to 10. The training set contains 50 images at the end of the interactive process. The classification performances are then provided for learning systems trained with only 10% of the database. In that case, the relative performances are more interesting than the absolute ones.

Quantitative evaluation for all the categories of ANN are summarized in table 2 and table 3, where the bp and MAP measures have been respectively reported.

category	M1	M2	M3	RETIN
arborgreens	58	73	72	**79**
campusinfall	56	65	70	**80**
cannonbeach	71	**86**	72	79
cherries	65	84	73	**86**
yellowstone	51	60	62	**63**
football	95	96	95	**100**
geneva	68	84	**94**	91
greenlake	56	63	67	**71**
sanjuans	66	**79**	71	72
springflowers	74	78	81	**86**
swissmountain	85	91	89	**95**

Table 2: ANN evaluation: system performances estimated with the breakeven metric bp ($\%$), at the end of the interactive learning process.

The RETIN strategy gives the best results for 8 categories out of 11 according to the bp measure, for 7 categories out of 11 according to the MAP measure, but the Bayesian M2 method is sometimes better, often close. The active M3 strategy provides poor results. The active learning seems to be very dependent on the number of training data; when this number is very small (only 50 here), the performances are poor. This observation joins the Tong's conclusion [30] about his technique. Our active strategy coupled with exploration steps is less sensitive and can succeed in task retrieval even when the training data set is very small.

category	M1	M2	M3	RETIN
arborgreens	63	81	79	**84**
campusinfall	65	72	81	**87**
cannonbeach	79	**91**	79	84
cherries	72	**94**	82	92
yellowstone	56	66	70	**73**
football	99	99	99	**100**
geneva	74	86	**98**	96
greenlake	60	66	75	**78**
sanjuans	74	**85**	79	80
springflowers	80	83	84	**91**
swissmountain	91	92	95	**98**

Table 3: ANN evaluation: system performances estimated with the MAP metric ($\%$), at the end of the interactive learning process.

6.3.2 COREL

Experiments on COREL are very interesting because the database is quite large, with many kinds of categories. In this context, comparison between systems to retrieve large and complex sets of images is meaningful.

The number m of images labeled at each feedback step is $m = 20$ and the number of feedbacks is 10. The training set contains 200 images at the end of the interactive learning process. The classification performances are then provided for systems trained with only 3% of the whole database.

First, we provide P/R curves on (respectively) doors (Fig. 3), flowers (Fig. 4) and landscape (Fig. 5) categories to illustrate the behavior of the methods on (respectively) an easy, a medium and a difficult category.

Most of the time, The RETIN strategy provides the best curves. Active learning strategies improve performances on the difficult retrieval task (Fig. 5), but RETIN is better than the other active strategy (SVM_{active} M3) on the 3 tested categories.

The bp values are reported in table 4, and MAP in table 5 for all the configurations.

Performances deeply depend on the complexity of the searched category. RETIN provides the best results for both bp and MAP statistics for all the categories. The number of exploration steps depends on the number of retrieved images but we noticed that it is quite stable. Active learning strategies improve performances, even if RETIN

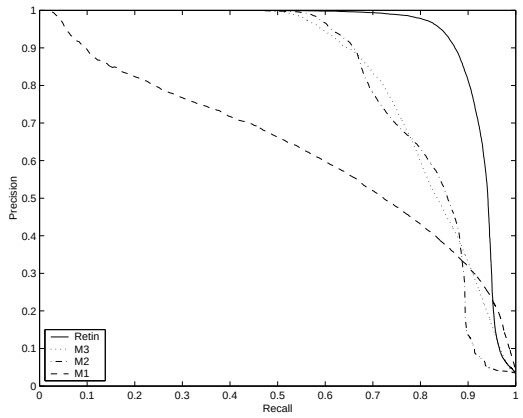

Figure 3: P/R curve for the doors category (COREL database).

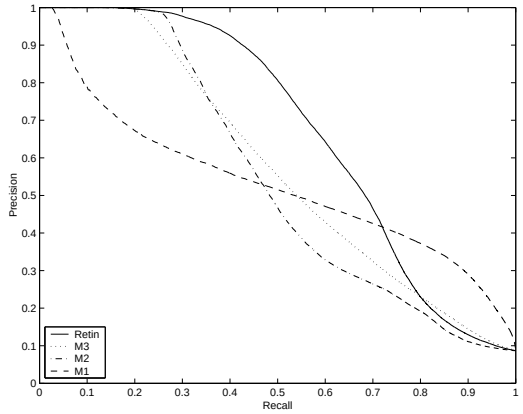

Figure 4: P/R curve for the flowers category (COREL database).

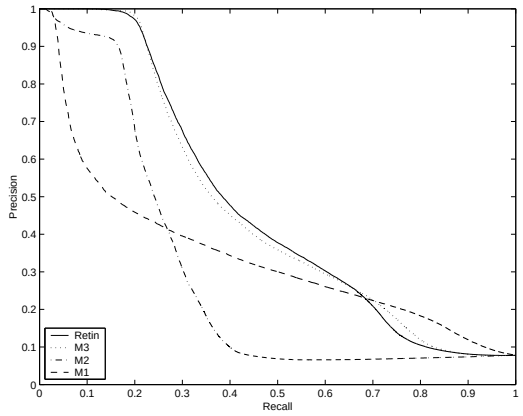

Figure 5: P/R curve for the landscape category (COREL database).

category	M1	M2	M3	RETIN
caverns	42	61	**70**	**70**
doors	60	73	75	**88**
flowers	51	49	52	**62**
savanna	34	32	39	**42**
landscape	36	30	43	**44**
Mean	45	49	56	**61**

Table 4: COREL evaluation: system performances estimated with the breakeven metric bp ($\%$), at the end of the interactive learning process.

category	M1	M2	M3	RETIN
caverns	40	62	**75**	**75**
doors	63	81	82	**93**
flowers	53	55	58	**67**
savanna	32	31	43	**45**
landscape	34	29	**47**	**47**
Mean	44	52	61	**66**

Table 5: COREL evaluation: system performances estimated with the MAP metric ($\%$), at the end of the interactive learning process.

always gives the best scores, SVM_{active} M3 strategy provides good results. For the most difficult category, the landscape category, both active techniques, M3 and RETIN, have the same performances. The exploration step seems to be helpless to boost the retrieval in that case. Last rows in tables 4 and 5 provide average performances. The RETIN strategy outperforms other techniques from 5% to 20%, which is a significant improvement in image retrieval context.

One can notice that the M1 method without active learning gives, most of the time, the worst results.

6.4 Computational aspects

The main computational needs is the $O(n)$ computation of membership to the relevant class (function f) on the whole database. Other requirements are negligible against n. In particular, the SVM optimization is not time consuming as soon as the number of training data (user labels) is small regarding n. In our experiments, all methods need about one second to be computed with a Pentium 3 GHz. With a one million image database, a similar configuration would require about 10 minutes to be computed.

7 Conclusion

In this article, we have presented an efficient interactive strategy for content-based image retrieval. The method is based on a two-step sequential algorithm with an exploration step followed by an active classification step.

The exploration step aims at providing a useful initial training data set for the next step of classification. This process is based on a stochastic sampling scheme. A multinomial law with simple and powerful settings of parameters is introduced in order to efficiently sample new images to display. In a few interaction iterations, the method provides a semantic query composed of all the images labeled as relevant by the user. Our strategy catches all the aspects of the semantic category in order to build a learning set of the searched category as various as possible.

For the classification task, we adapted a SVM classifier to CBIR context. We also introduced an active learning strategy to select for labeling new images close to the boundary between relevant and irrelevant images. This method allows to get good performances of classification with few training sets. This is definitively a major advantage in CBIR context where the user interaction has to be as weak as possible.

The method has been validated through experiments on large databases with specific grouping of images to get complex categories. We implemented leader active learning methods and a Bayesian classification for comparison. Our two-step strategy outperforms other techniques from 5% to 20% on the COREL database. In image category retrieval, the two steps complement very well each other: the first step aims at retrieving images from several modes scattered in the feature space, while the second step efficiently determines the boundary of the category.

Our currently works deal with the evaluation of the scalability of theses techniques when huge databases are considered. We are convinced that, for category search in very large databases, efficient exploration process before classification process will become crucial.

References

[1] R. Brunelli and O. Mich. Histograms analysis for image retrieval. *Pattern Recognition*, 34:1625–1637, 2001.

[2] G. Caenen, G. Frederix, A. Kuijk, E. Pauwels, and B. Schouten. Show me what you mean! PARISS: A CBIR-interface that learns by example. In *International Conference on Visual Information Systems (Visual'2000)*, volume 1929, pages 257–258, 2000.

[3] E. Chang, B. T. Li, G. Wu, and K. Goh. Statistical learning for effective visual information retrieval. In *IEEE International Conference on Image Processing*, Barcelona, September 2003.

[4] E. Chang, B. T. Li, G. Wu, and K. Goh. Statistical learning for effective visual information retrieval. In *IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003.

[5] O. Chapelle, P. Haffner, and V. Vapnik. Svms for histogram based image classification. *IEEE Transactions on Neural Networks*, 9, 1999.

[6] D. Cohn. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[7] M. Cord, J. Fournier, and S. Philipp-Foliguet. Exploration and search-by-similarity in CBIR. In *IEEE 15th Sibgrapi (Symp. on Comp. Graphics and Im. Processing)*, Sao Carlos, Brazil, October 12 - 15 2003.

[8] I. Cox, M. Miller, T. Minka, T. Papathomas, and P. Yianilos. The bayesian image retrieval system, PicHunter: Theory, implementation and psychophysical experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, 2000.

[9] J. Eakins. Towards intelligent image retrieval. *Pattern Recognition*, 35:3–14, 2002.

[10] J. Fournier and M. Cord. Long-term similarity learning in content-based image retrieval. In *IEEE International Conference in Image Processing (ICIP'02)*, Rochester, New-York, USA, September 2002.

[11] J. Fournier, M. Cord, and S. Philipp-Foliguet. Retin: A content-based image indexing and retrieval system. *Pattern Analysis and Applications Journal, Special issue on image indexation*, 4(2/3):153–173, 2001.

[12] D. Geman and R. Moquet. A stochastic feedback model for image retrieval. In *RFIA'2000*, volume III, pages 173–180, Paris, France, February 2000.

[13] P. Gosselin and M. Cord. A comparison of active classification methods for content-based image retrieval. In *International Workshop on Computer Vision meets Databases (CVDB), ACM Sigmod*, pages 51–58, Paris, France, June 2004.

[14] P. Gosselin, M. Najjar, M. Cord, and C. Ambroise. Discriminative classification vs modeling methods in CBIR. In *IEEE Advanced Concepts for Intelligent Vision Systems (ACIVS)*, Brussel, Belgium, September 2004.

[15] T. Hastie, R. Tibshirani, and J. Friedman. *The Element of Statistical Learning*. Springer, 2001.

[16] Y. Ishikawa, R. Subramanya, and C. Faloutsos. MindReader: Query databases through multiple examples. In *24th VLDB Conference*, pages 218–227, New York, 1998.

[17] T. Joachims. Transductive inference for text classification using support vector machines. In *Proc. 16th International Conference on Machine Learning*, pages 200–209. Morgan Kaufmann, San Francisco, CA, 1999.

[18] D. Lewis and J. Catlett. Heterogenous uncertainty sampling for supervised learning. In *International Confenrence on Machine Learning*, 1994.

[19] A. Mojsilovic and B. Rogowitz. Capturing image semantics with low-level descriptors. In *International Conference in Image Processing (ICIP'01)*, volume 1, pages 18–21, Thessaloniki, Greece, October 2001.

[20] N. Najjar, J. Cocquerez, and C. Ambroise. Feature selection for semi supervised learning applied to image retrieval. In *IEEE ICIP*, Barcelena, Spain, Sept. 2003.

[21] J. Park. On-line learning by active sampling using orthogonal decision support vectors. In *IEEE Neural Networks for Signal Processing*, 2000.

[22] G. Patanè and M. Russo. The enhanced LBG algorithm. *IEEE Transactions on Neural Networks*, 14(9):1219–1237, November 2001.

[23] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *International Confenrence on Machine Learning*, 2001.

[24] Y. Rui and T. Huang. Optimizing learning in image retrieval. In *Conf on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 236–243, Hilton Head, SC, June 2000.

[25] Y. Rui, T. Huang, S. Mehrotra, and M. Ortega. A relevance feedback architecture for content-based multimedia information retrieval systems. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 92–89, 1997.

[26] S. Santini, A. Gupta, and R. Jain. Emergent semantics through interaction in image databases. *IEEE Transactions on Knowledge and Data Engineering*, 13(3):337–351, 2001.

[27] C. Schmid. Weakly supervised learning of visual models and its application to content-based retrieval. *International Journal of Computer Vision*, 56(1-2):7–16, January,February 2004.

[28] B. Scholkopf. Estimating the support of high-dimensional distribution. Technical report, Microsoft Reseach, 1999.

[29] A. Smola and B. Scholkopf]. *Learning with kernels*. MIT Press, Cambridge, MA., 2002.

[30] S. Tong. *Active Learning: Theory and Applications*. PhD thesis, Stanford University, 2001.

[31] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *ACM Multimedia*, 2001.

[32] L. Tran and R. Lenz. PCA-based representation of color distributions for color-based image retrieval. In *International Conference in Image Processing (ICIP'01)*, volume 2, pages 697–700, Thessaloniki, Greece, October 2001.

[33] T. Tuytelaars and L. V. Gool. Content-based image retrieval based on local affinely invariant regions. In *Third Int'l Conf. on Visual Information Systems, Visual99*, pages 493–500, 1999.

[34] V. Vapnik. *Statistical Learning Theory*. John Wiley, New York, 1998.

[35] N. Vasconcelos. *Bayesian models for visual information retrieval*. PhD thesis, Massachusetts Institute of Technology, 2000.

[36] N. Vasconcelos and M. Kunt. Content-based retrieval from image databases: current solutions and future directions. In *International Conference in Image Processing (ICIP'01)*, volume 3, pages 6–9, Thessaloniki, Greece, October 2001.

[37] R. Veltkamp. Content-based image retrieval system: A survey. Technical report, University of Utrecht, 2002.

[38] L. Wang. Image retrieval with svm active learning embedding euclidian search. In *IEEE International Conference on Image Processing*, Barcelona, September 2003.

[39] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, 2003.