Graph-Cut Transducers for Relevance Feedback in Content Based Image Retrieval[1]

Hichem Sahbi

Jean Yves Audibert

Renaud Keriven

Research Report 07-30
February 2007

CERTIS, ENPC,
77455 Marne la Vallee, France,

Graph-Cut Transducers for Relevance Feedback in Content Based Image Retrieval[2]

Coupes Minimales pour le Controle de Pertinence en Recherche d'Images

*

Abstract

Closing the semantic gap in content based image retrieval (CBIR) basically requires the knowledge of the user's intention which is usually translated into a sequence of questions and answers (Q&A). The user's feedback to these questions provides a CBIR system with a partial labeling of the data and makes it possible to iteratively refine a decision rule on the unlabeled data. Training of this decision rule is referred to as transductive learning.

This work is an original approach to relevance feedback (RF) based on graph-cuts. Training consists in implicitly modeling the manifold enclosing both the labeled and unlabeled dataset and finding a partition of this manifold using a min-cut. This RF model exploits the structure of the manifold by considering also the structure of the unlabeled data. Experiments conducted on generic as well as specific databases show that our graph-cut based approach is very effective, outperforms other existing methods and makes it possible to converge to almost all the images of the user's "class of interest" with a very small labeling effort.

Table des matières

1   Introduction

Two interrogation modes are known in CBIR ; the query by example and relevance feedback. In the first mode the user submits a query image as an example of his "class of interest" and the system displays the closest image(s) using a feature space and a suitable metric [2, 30]. In the second category (see the pioneering works [19, 23, 26]) the user labels a subset of images as positive and/or negative according to an unknown metric defined in "his mind" and the CBIR system refines a metric and/or a decision rule and displays another set of images hopefully closing the gap between the user's intention and the response(s) of the the CBIR system [36, 8, 25, 37]. This process is repeated until the system converges to the user's class of interest. The performance of an RF system is usually measured as the expectation of the number of user's responses (or iterations) necessary to focus on the targeted class. This performance depends on the capacity of an RF system (i) to generalize well on the set of unlabeled images using the labeled ones, (ii) to ask the most informative questions to the user (see for instance [32]) and (iii) the consistency and self-consistency of the user's responses. Points (i)–(ii) are respectively referred to as *the transduction* and the *display models*. Point (iii) assumes that different users have statistically the same answers according to an existing but unknown model referred to as the *user model*.

Different schemes exist in the literature for the purpose of RF [35, 25, 37] which are either based on density estimation [21, 14] or discriminative training [32], depending respectively on the fact that they model the distribution of the positive and (possibly) the negative labeled images or they build a decision function which classifies the unlabeled data. In the first category, different density estimation methods are used in RF including non parametric Parzen windows [21], Gaussian mixture models [8], logistic regression [4] and novelty detectors [5, 29]. In [6, 7], the authors introduced a notion of relative judgment of the user, i.e., the response is not binary but a relative number measuring the relevance of a displayed set of images. The user's response is assumed as a sigmoid function of the distance, so images close to the highly numbered set are more likely to be the target than the others. The authors used Gaussian mixture models and a Bayesian framework in order to estimate (and update) a distribution though all images and display those which the highest probability. In [13, 9], the authors defined a new and original RF model which considers that the targeted class is a mental picture. The proposed approach defines a criteria based on the mutual information between the user's responses and all the possible target images in the database and display those which maximize this criteria.

In the second family, discriminative methods learn from the aggregated set of

positive and negative labeled images how to classify the unlabeled ones. Existing RF methods use support vector machines [32, 32, 10], decision trees [20], boosting [31] and Bayesian classifiers [10, 34, 7, 6]. The RF method in [32] shows a particular interest by its important gain in the convergence speed when using active learning [28, 1].

In this paper, we introduce a new RF scheme based on graph-cuts [3]. Our RF approach builds a decision boundary using a subset of images labeled (positively and negatively) by the user. We first model the topology of the image set, including the unlabeled images, using a graph, then we partition this graph using the min-cut method [3]. This is strictly equivalent to minimizing an energy function containing (1) a fidelity term ensuring the consistency of the labels of the graph partition with those provided by the user and (2) a regularization term ensuring that neighboring data are likely to have the same labels. In contrast to existing methods which only rely on the labeled set of images, our approach integrates the unlabeled data. These unlabeled data turn out to be very useful when only few labeled data are available since it allows to favor decision boundaries located in low density region of the image space, which are very often encountered in practice.
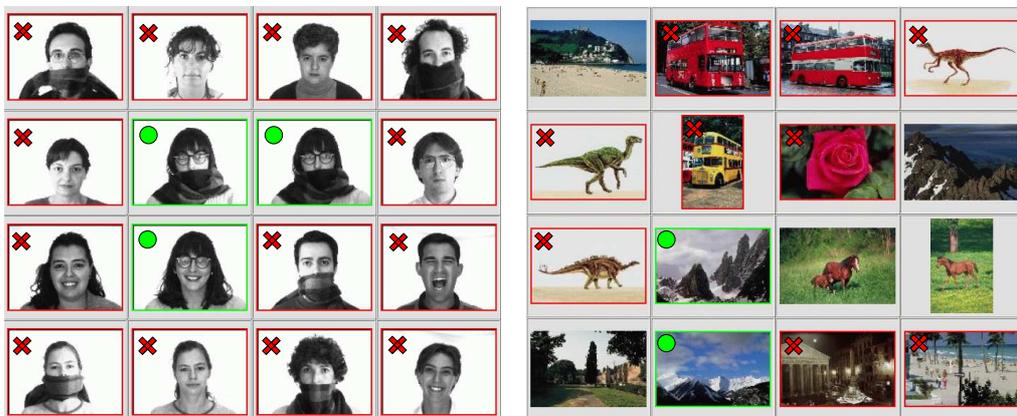

FIG. 1 − An example of an RF display, where the user's response is unique on the ARF database (left) and non unique on the Corel dataset (right) ; when searching for the "nature" category, it is clear that the labeling is not unique. Red crosses stand for the negative labeled images while the green disks stand for the positive ones (No color stands for no labels).

In the reminder of this paper, we consider the following notation. X is a random variable standing for a training sample taken from $\mathcal{X}$ and Y its class label in $\{+1, -1\}$ ($Y = 1$ if the sample X belongs to the targeted class and -1 otherwise). $G = \langle V, E \rangle$ denotes a graph where V is a set of vertices and E are weighted

edges. We use also k, t as indices for iterations. Among terminologies a *display* is a set of images taken from the database which are shown to the user at iteration t. The paper is organized as follows : Section 2 introduces the overall architecture of the RF process. Section 3 describes our RF model based on graph-cuts and the different strategies for the display model. Section 4 provides an extensive experimental study using different databases including specific ones ; leaf and face databases. We discuss the method and we conclude in Section 5 while providing a direction for a future work.

2   The Interaction Model

Denote $\mathcal{S} = \{X_1, ..., X_n\}$, $\mathcal{Y} = \{Y_1, ..., Y_n\}$ respectively as a training set of images and the underlying unknown ground truth. Here Y_i is equal $+1$ if the image X_i belongs to the user's "class of interest" and $Y_i = -1$ otherwise. Let us consider $\mathcal{D}_t \subset \mathcal{S}$ as a display shown at iteration t and $\mathcal{Y}_t$ the labels of $\mathcal{D}_t$. Our interaction consists in asking the user questions such that his/her responses make it possible to reduce the *semantic gap* according to the following steps :

- "Page Zero" : Select a display $\mathcal{D}_1$ which might be a random set of images or the prototypes found after applying clustering or Voronoi subdivision [24].
- Reduce the "semantic gap" iteratively ($t = 1,..., T$) :

  1. Label the set $\mathcal{D}_t$ using a (possibly stochastic) *known-only-by-the-user* function denoted $\mathcal{L}$,

  $$\mathcal{Y}_t \ \leftarrow \ \mathcal{L}\left(\mathcal{D}_t\right) \tag{1}$$

  2. Train a decision function $f_t : \mathcal{X} \rightarrow \{-1, +1\}$ on the (so far) labeled training set $\mathcal{T}_t = \bigcup_{k=1}^{t}(\mathcal{D}_k, \mathcal{Y}_k)$.
  3. Select the next display $\mathcal{D}_{t+1} \subset \mathcal{S} - \bigcup_{k=1}^{t} \mathcal{D}_k$.

2.1   Transduction Model

Training consists in finding a decision function in order to classify the unlabeled data. The transduction model is the one used for this training. At iteration t, we only know the labeled training set $\mathcal{T}_t$ and the unlabeled set of images $\mathcal{S} - \cup_{k=1}^{t}\mathcal{D}_k$. The target of the transduction model is to use efficiently these data to estimate the actual decision function

$$\underset{f:\mathcal{X}\rightarrow\{+1;-1\}}{\operatorname{argmin}} \ P\left[f\left(X\right) \ \neq \ Y\right].$$

In our setting, it is important to generalize well even when the size of the labeled training set is small. This is why this step should use transductive methods which implicitly assume that the topology of the decision boundary depends on the unlabeled set $\mathcal{S} - \cup_{k=1}^{t} \mathcal{D}_k$. More precisely, the clustering assumption implicitly made is : the decision boundary is likely to be in low density regions of the input space $\mathcal{X}$ [22].

2.2   Display Model

The convergence of the RF model to the actual decision boundary is very dependent on the amount of information provided by the user. The display model is a sampling strategy which selects a collection of images that improves our current estimate of the "class of interest". This can be achieved by showing samples of difficult-to-classify images such as those close to the decision boundary. Given the labeled set $\mathcal{T}_t$, and let $f_{\mathcal{D}}$ be a classifier trained on $\mathcal{T}_t$ and a display $\mathcal{D}$. The issue of selecting $\mathcal{D}_{t+1}$ can be formulated at iteration $t+1$ as :

$$\mathcal{D}_{t+1} \leftarrow \underset{\mathcal{D}}{\operatorname{argmin}} \quad P\left[f_{\mathcal{D}}(X) \neq Y\right]$$

$$s.t. \quad \mathcal{D}_{t+1} \bigcap \left(\bigcup_{k=1}^{t} \mathcal{D}_k\right) = \emptyset$$

(2)

The constraint ensures disjoint displays. In practice, finding $\mathcal{D}_{t+1}$ among all the possible subsets in $\mathcal{S} - \bigcup_{k=1}^{t} \mathcal{D}_k$ is clearly intractable as $P(.)$ is unknown and the the whole process is computationally expensive. Heuristics have to be used in order to explore the space of display hypotheses.

2.3   User Model

The user model is defined as the mapping function $\mathcal{L}$ which, given a display $\mathcal{D}_t$, provides the labels $\mathcal{Y}_t$. When the ground truth is unique, this function is consistent (through different users) and self-consistent (with respect to the same user) so the user's answer is coherent and objective, otherwise the labeling function becomes stochastic and dependent on different classes of users. The coherence is defined as the probability (given a targeted class and a display) that the user's response is unique (see Figure 1). In this work we only consider consistent and self-consistent users.
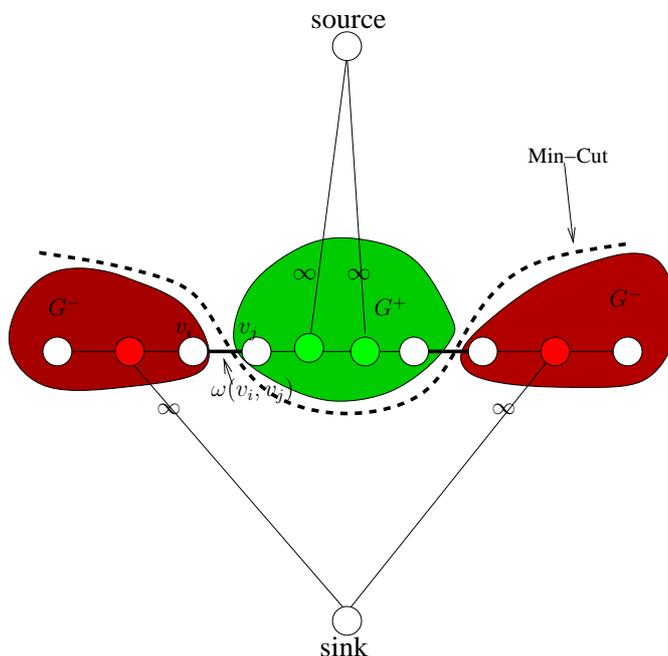
FIG. 2 – This figure shows a didactic $1D$ graph example and its min-cut. Green (resp. red) colored nodes correspond to the positive (resp. negative) labeled images.

3   Graph-cuts and Relevance Feedback

Considering the manifold enclosing the training set $\mathcal{S}$, we turn the training problem into a segmentation one. Using a variational framework, we design some energy function that should be minimized by the desired segmentation. Our energy is compatible with the graph-cut method[3] so that a global minimum can be computed very efficiently.

3.1   The Energy Function

Let us fix the iteration number t and consider the set $\mathcal{S}$ in which only a subset $\bigcup_{k=1}^{t} \mathcal{D}_k \subset \mathcal{S}$ is labeled. Training consists in finding a decision function f_t and a configuration of the labels $\mathcal{Y}$ of $\mathcal{S}$ that minimizes the following energy function :

$$\mathcal{E}(\mathcal{S}, \mathcal{Y}) \;=\; \sum_{i=1}^{n} D_i\left(Y_i\right) \,+\, \lambda \sum_{i=1}^{n} \sum_{X_j \in \mathcal{N}_i} V_{ij}\left(Y_i, Y_j\right)$$
$$Y_i \;\in\; \{-1, +1\}, \quad i = 1, ..., n \tag{3}$$

The first term, referred to as the fidelity term, measures the confidence when labeling the training sample X_i as Y_i while the second one is a regularizer which

ensures that training samples in the neighborhood of X_i (denoted $\mathcal{N}_i$) are assigned the same (or close) labels as X_i. The positive parameter λ controls the trade-off between fidelity and regularization.

In practice, we choose :

$$D_i\left(Y_i\right) = \begin{cases} +\infty & \text{if } X_i \in \cup_{k=1}^{t}\mathcal{D}_k \text{ and } Y_i = \mathcal{L}(X_i) \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

so that labeled data $\bigcup_{k=1}^{t} \mathcal{D}_k$ are assigned infinite confidence while :

$$V_{ij}(Y_i, Y_j) = \mathbb{1}_{\{Y_i \neq Y_j\}} \times w_{ij} \tag{5}$$

Here w_{ij} captures the similarity between X_i and X_j (see Section 3.2) and $\mathbb{1}_{\{\}}$ is the indicator function. The energy $\mathcal{E}$ is minimized using graph-cuts [3]. This is actually possible thanks to the submodularity of our regularizing term (i.e. $V_{ij}(-1, -1) + V_{ij}(1, 1) \leq V_{ij}(-1, 1) + V_{ij}(1, -1)$, see [18]). Indeed, we use the classical generalized Potts model.

Let us just sketch the principles of this method. We construct a graph $G = \langle V, E \rangle$. Nodes v_i represent the X_is, while two terminal nodes s (the source) and t (the sink) are added. Between two non terminal nodes, the edges are symmetric with weight $w(v_i, v_j) = w(v_j, v_i) = w_{ij}$. Terminal links are added with the following weights : $w(s, v_i) = D_i(+1)$ and $w(v_i, t) = D_i(-1)$. Looking for a global optimum of the energy function is equivalent to finding an st-cut of G of minimal cost. This cut partitions the graph G into two disconnected subgraphs denoted G^+ and G^- (see Figure 2) such that (1) $s \in G^+$, (2) $t \in G^-$, and (3) a sample X_i is assigned to the positive (resp. negative) class if the underlying node v_i belongs to G^+ (resp. G^-). Finding the minimal cut is in turn equivalent to maximizing the flow from s to t in G [12] and many fast algorithms are available to find this flow, among which the *graph-cut* method proposed in [3]. Millions of nodes are handled without difficulties by the original algorithm. Yet, if efficiency becomes an issue, recent extensions like dynamic graph-cuts[17] or active graph-cuts[15] are available for incremental problems like ours.

3.2 Weighting

A link $e_{ij} \in E$ is weighted using a correlation function measuring the similarity of X_i and X_j. The Gaussian similarity function is :

$$w_{ij} \;\; = \;\; \exp\left(-\frac{\|X_i - X_j\|^2}{2\,\sigma^2}\right) \tag{6}$$

This weight defines an inner product in a high dimensional mapping space i.e., there exists a mapping from the input space into an infinite dimensional Hilbert space such that : $w_{ij} = \langle \Phi(X_i), \Phi(X_j) \rangle$ also denoted $k(X_i, X_j)$. The choice of the scale parameter σ is crucial and in practice it is found by cross validation. When σ is underestimated, i.e., $\sigma \to 0$, $k(X_i, X_j) \to 0$, so all the edges in G, except the s and t links, will be assigned the same weights (i.e. 0) and any cut which shatters G into two sub-graphs G^+ and G^- (such that $s \in G^+$ and $t \in G^-$) will be a min-cut. Clearly, the min-cut solution and labeling of the graph is not unique and unstable. On the other hand an overestimated σ, i.e., $\sigma \to \infty$, makes $k(X_i, X_j)$ close to 1 so the number of edges intervening in a min-cut is minimized (and not the similarity). When training data show large variations in scale it is difficult to find an appropriate setting for this scale parameter (see [11]).

Triangular Kernel : this function introduced in [11] is defined as :

$$w_{ij} \quad = \quad k_T(X_i, X_j) = -\|X_i - X_j\| \tag{7}$$

It is clear that $V(Y_i, Y_j) = \mathbb{1}_{\{Y_i \neq Y_j\}} k_T(X_i, X_j)$ is sub-modular and that using $k_T(X_i, X_j)$, the solution minimizing (3) is rotation and translation invariant. We will now show that the solution is also scale invariant. For this, let $\mathcal{S}^\gamma = \{\gamma X_1, ..., \gamma X_n\}$ be a scaled version of $\mathcal{S}$. We have

$$
\begin{aligned}
\mathcal{E}(\mathcal{S}^\gamma, \mathcal{Y}) \;=\; & \sum_{i=1}^{n} D_i\,(Y_i) \\
& + \sum_{i=1}^{n} \sum_{X_j \in \mathcal{N}_i} -\mathbb{1}_{\{Y_i \neq Y_j\}} \|\gamma\, X_i \;-\; \gamma\, X_j\|
\end{aligned}
$$

From (4), we have $D_i\,(Y_i) = \gamma D_i\,(Y_i), \forall\, \gamma$, hence :

$$
\begin{aligned}
\mathcal{E}(\mathcal{S}^\gamma, \mathcal{Y}) \;=\; & \gamma \sum_{i=1}^{n} D_i\,(Y_i) \\
& + \gamma \sum_{i=1}^{n} \sum_{X_j \in \mathcal{N}_i} -\mathbb{1}_{\{Y_i \neq Y_j\}} \|X_i - X_j\| \\
\;=\; & \gamma\, \mathcal{E}(\mathcal{S}, \mathcal{Y})
\end{aligned}
$$

which also results from the fact that the neighbors $\mathcal{N}_i$ of X_i are invariant to a global scaling of $\mathcal{S}$. Consequently, $\arg \min_{\mathcal{Y}} \mathcal{E}(\mathcal{S}^\gamma, \mathcal{Y}) = \arg \min_{\mathcal{Y}} \mathcal{E}(\mathcal{S}, \mathcal{Y})$.
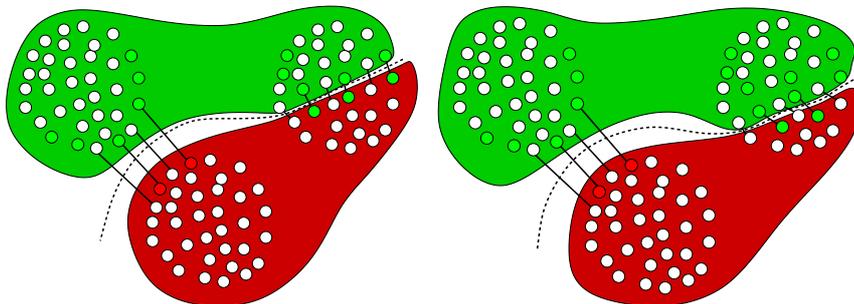
FIG. 3 – An example of the evolution of the min-cut when using iteratively the *decision boundary refinement* strategy. Data displayed to the user corresponds to the min-cut nodes and the user's responses (corrections) are shown in red and green. The image in the left hand side shows the user's correction on the negative class. The image in the right hand side shows the movement of the decision boundary when the user actively corrects the labels of the min-cut data.

3.3 Display strategies

The goal of a display strategy, at a given iteration t, is to select a display $\mathcal{D}_{t+1} \subset \mathcal{S} - \bigcup_{k=1}^{t} \mathcal{D}_k$ minimizing the generalization error of the next prediction f_{t+1} of the decision function, which will hopefully converge to the actual decision function. A brute search strategy which consists in exploring all possible displays $\mathcal{D}$ in $\mathcal{S} - \bigcup_{k}^{t} \mathcal{D}_k$, building a decision function f on $(\bigcup_{k}^{t} \mathcal{D}_k) \bigcup \mathcal{D}$ and estimating their generalization errors is *clearly* out of hand. In practice, we have to consider display heuristics.

If our inference method does not use the unlabeled data, i.e. if we use an inductive learner and not a transductive one, then the display heuristics are the one of active learning. Research in this field tells us that one should be cautious in using these heuristics since many of them, which has nevertheless led to advances in numerous application fields, can perform worse than the basic display strategy consisting in choosing uniformly randomly the images of the display (see [1] and references within for a more detailed discussion).

When we use the unlabeled data by using a transductive algorithm, the heuristics still rely on the following basic notions : at each iteration, one can select the display in order to refine the current estimate of the decision boundary or one can select the display in order to find uncharted territories in which the actual decision boundary is present. The first display strategy *exploits* our knowledge of the likely position of the decision boundary while the second one *explores* new regions. We believe that any good CBIR system should find the correct balance between ex-

ploration and exploitation).

In our work, we use three display strategies which either explore the different modes of the user's "class of interest" and/or refine locally each mode.

1. *Exploitation or Decision Boundary Refinement :* the display $\mathcal{D}_{t+1}$ corresponds to the unlabeled nodes of the min-cut edges in G (see Figure 3). This strategy is efficient when the user's "class of interest" contains a single mode, i.e. when the part of the input space that should be labeled "class of interest" is topologically connected. The goal is to display ambiguous images, close to the boundary of the "class of interest", which might be misclassified by f_t. Figure 3 shows an example of the evolution of the decision boundary where at each iteration only images belonging to the min-cut edges are shown to the user.

2. *Exploration or Mode Search :* the display $\mathcal{D}_{t+1}$ is randomly selected among unlabeled samples far from the decision boundary.

3. *Combined Exploration and Exploitation :* the display $\mathcal{D}_{t+1}$ is selected by combining the two above strategies.
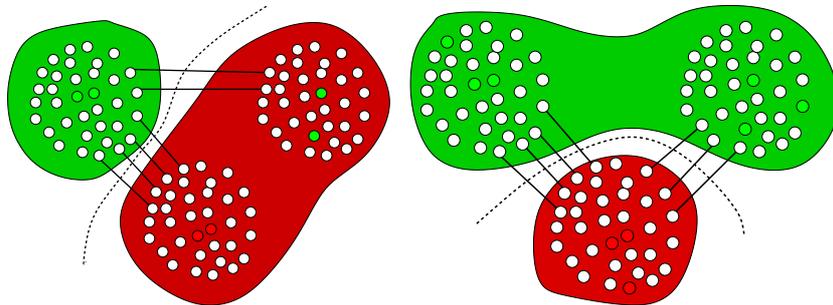

FIG. 4 – An example of the evolution of the min-cut when using iteratively the *mode search* strategy. The user' responses are shown in red and green. The image in the left hand side shows the user's correction on the negative class. The image in the right hand side shows the movement of the decision boundary when the user actively corrects the labels of randomly displayed data.

4 Experiments

In this section, we demonstrate the validity of relevance feedback using the graph-cut transducers. We compare it to popular state-of-the-art methods inclu-

ding support vector machines and Bayesian inference. The effectiveness is measured as the expected number of images per class which are displayed to the user or equivalently the average number of iterations necessary in order to show a fraction of images per class.

4.1 Databases

Experiments were conducted on simple databases (Olivetti and Swedish) as well as difficult ones (Corel) in order to show both the good performance of our RF scheme and the improvement brought by graph-cut transducers with respect to existing state-of-the-art models. The Olivetti face database contains 40 persons each one represented by 10 faces. Each face is processed using histogram equalization and encoded using kernel principal component analysis (KPCA) [16] resulting into 20 coefficients. The Swedish set contains 15 categories of leaf silhouettes each one represented by 75 contours. Each contour $\mathcal{C}$ is encoded using 14 coefficients corresponding to the eigenvalues of KPCA on $\mathcal{C}$ [27]. The Corel database contains 10 categories each one represented by 100 images. This database is generic and images rang from simple objects to natural scenes with complex background (see Figure 8). Each image in this database is encoded simply using a $3D$ RGB color histogram of 125 dimensions. Notice that this simple histogram signature makes classes very spread so the relevance feedback task is more challenging.

Notice that for all these databases the ground truth is provided. Given a display $\mathcal{D}_t$ and a user, we assume that the labeling function $\mathcal{L}$ is unique so we can perform simulations.

4.2 Benchmarking

We evaluate the performance of our RF scheme using two measures, recall and generalization.
Recall : let Z_t be a random variable standing for the total number of relevant images returned by the CBIR system until iteration t, i.e., those belonging to the user's "class of interest". Let K be the cardinality of the classes of interest, the recall is defined as :

$$E(Z_t) \ = \ \sum_{r=1}^{K} r \ P(Z_t = r),$$

here the randomness and the expectation of Z_t is taken through different classes of interest.

Balanced generalization error : we also measure the performance of our RF system by the empirical error of the classifier f_t at a given iteration t. This measure is given by :

$$\sum_{i=1}^{n} \frac{1}{2n_+} \mathbb{1}_{\{f_t(X_i) \neq \mathcal{L}(X_i)=1\}} + \sum_{i=1}^{n} \frac{1}{2n_-} \mathbb{1}_{\{f_t(X_i) \neq \mathcal{L}(X_i)=-1\}},$$

where $n_+ = \# \{X_i, \mathcal{L}(X_i) = 1\}_{i=1}^{n}$ and $n_- = n - n_+$. Notice that the increase of the recall does not imply necessarily an increase in generalization and vice versa. The rational is that the former depends mainly on the display strategy. For instance if the generalization error of the graph-cut transducer decreases, and if we consider a "bad" display strategy (as random) then the recall, at a given iteration t, might not increase significantly (see Figure 6). Conversely, if the recall increases by exploring new orthant of the feature space, this may degrade the performance of the graph-cut transducer temporarily (for instance ; Figure 6, bottom shows an oscillation of the generalization error, on the Swedish set, even though the recall is strictly increasing). We choose to show both these two measures as they are not strictly correlated but complementary.

4.3   Settings

Different settings were experimented for our RF system including the size of the neighborhood (denoted NN) when building the graph G and the display strategies. The choice of these parameters has been experimented and will be discussed in this section for different test sets.

Topology : the parameter NN (the degree of the graph G) is very dependent on the topology of the manifold enclosing the training set (particularly on its curvature). In practice, we set different values of NN and we estimate the best recall and generalization performances of our RF system. The results show that the best performance are achieved when $NN = 4$ on both the Olivetti and the Swedish sets (see Figure 5).

Display strategies : we also investigated the impact of the display strategies on the performance of our RF system. Figure (6), shows the recall and the generalization error for different strategies. It appears that the local boundary refinement strategy outperforms the combined exploitation-exploration ; the rational is that the Olivetti and the Swedish sets contain homogeneous classes, so it makes more sense to refine the search locally than exploring new modes. In the case of Corel the results show that the combined strategy provides better results in terms of recall and generalization, as the modes of the "classes of interest" are spread. This
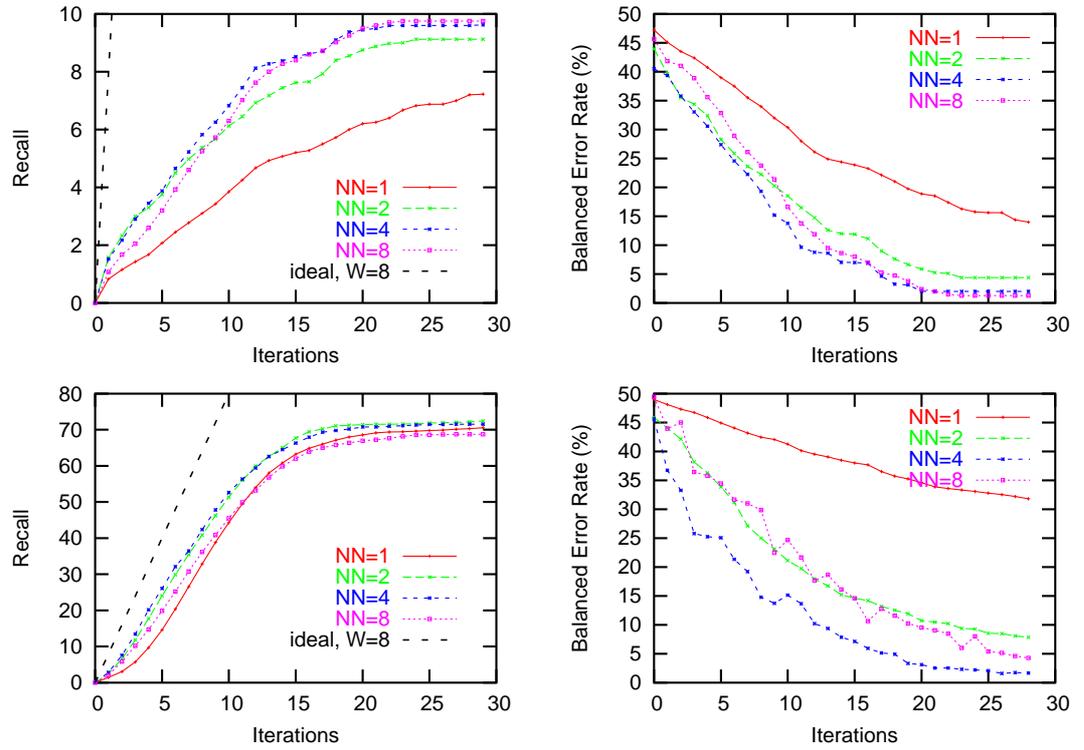
FIG. 5 – This picture shows the variation of the recall and generalization with respect to the number of neighbors in the graph ; using the Olivetti(top) and the Swedish(bottom). For both cases we reach the highest recall when $NN = 4$. We use the local boundary refinement strategy for the display.

improvement is brought mainly at the end of the interaction process ; when all the modes were explored (see Figure 6, bottom).

4.4   Comparison

We compared our method to standard representative relevance feedback tools including support vector machines (SVMs) and Bayesian inference (based on Parzen windows). In both SVMs and Parzen estimators we use exactly the same kernel function (i.e., triangular) and the same display strategy as graph-cuts (i.e., combined exploration exploitation). An extensive study in [10] showed that SVM based relevance feedback using the triangular kernel achieved far better results than other kernels, so we limit our comparison to SVM and Parzen using this kernel only. In what follows, we briefly remind the principle of relevance feedback using SVMs and Bayesian inference.

Bayesian Inference : given $X \in \mathcal{X}$, we assign X to the class decided by the sign of :

$$P(Y = 1|X) \ - \ P(Y = -1|X) \tag{8}$$

For a given set of aggregated labeled displays $\mathcal{T}_t$, we model the density $P(X|Y = 1)$ using the Parzen estimator as in [21] :

$$P(X|Y = 1) \ = \ \frac{1}{S|\mathcal{I}_t^+|} \ \sum_{i \in \mathcal{I}_t^+} k\left(\|X - X_i\|\right), \tag{9}$$

where k is the triangular kernel, $\mathcal{I}_t^+ = \{i : Y_i = 1\}$ and S is a normalizing constant depending only on k. Equivalently, we can model $P(X|Y = -1)$. Now, assuming the priors $P(Y = 1)$, $P(Y = -1)$ identical, the class of a given X is simply decided by the sign of

$$\frac{1}{|\mathcal{I}_t^+|} \sum_{i \in \mathcal{I}_t^+} k\left(\|X - X_i\|\right) - \frac{1}{|\mathcal{I}_t^-|} \sum_{i \in \mathcal{I}_t^-} k\left(\|X - X_i\|\right)$$

Support Vector Machines : SVM are reported to be one of the best tools in relevance feedback (see for instance [32]) as they generalize well even with few labeled samples. Following the same notation, and given $\mathcal{T}_t$, SVM training consists in (i) implicitly mapping training samples in $\mathcal{T}_t$ from the input space $\mathcal{X}$ into a high dimensional Hilbert space $\mathcal{H}$ via a function $\Phi()$ characterized by a kernel function k and (ii) finding a hyperplane in $\mathcal{H}$ which, in a loose sense, maximizes the margin between the positive set $\mathcal{T}_t^+ = \{(X_i, Y_i) : Y_i = +1\}$ and the negative
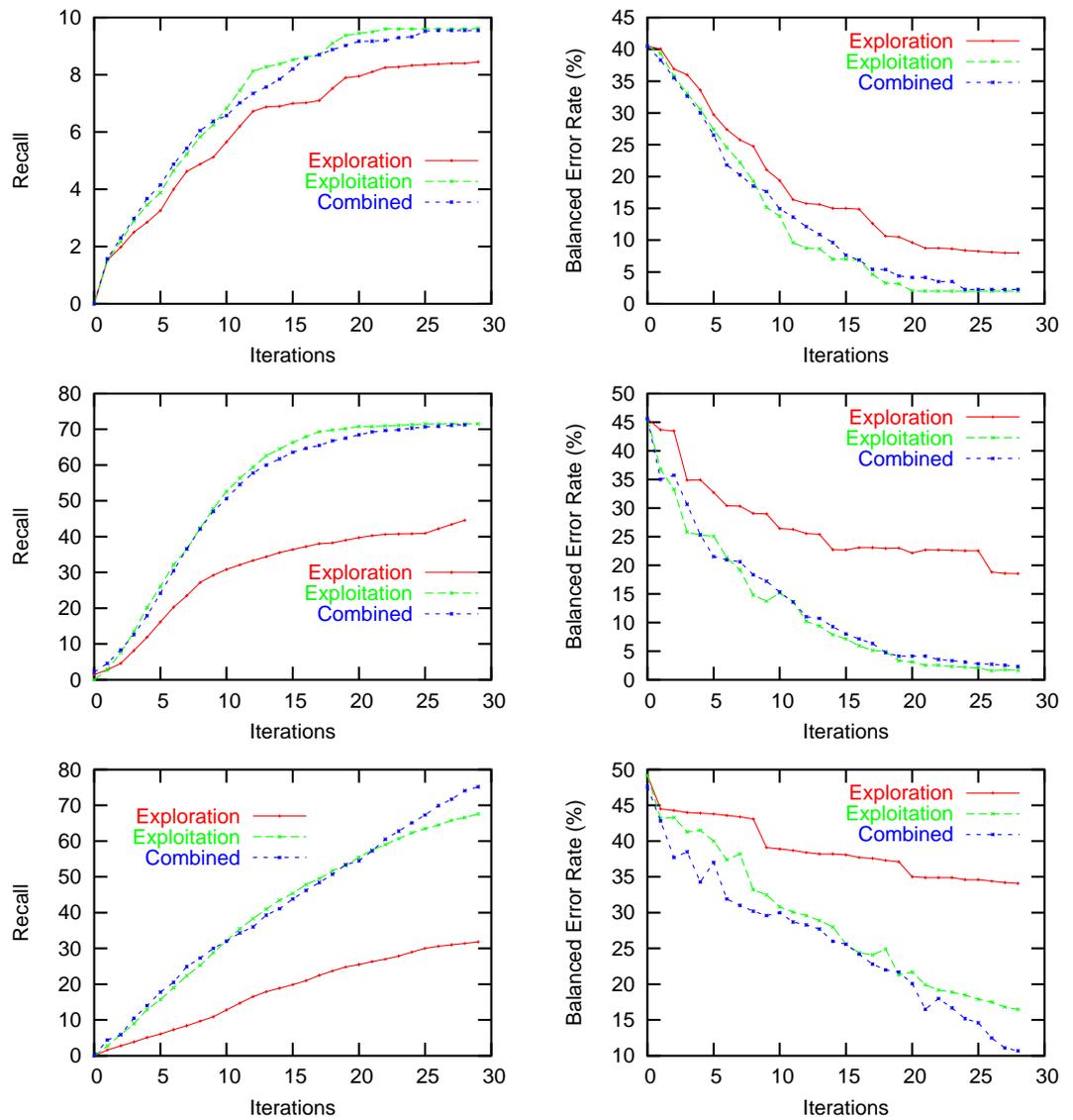
FIG. 6 – This picture shows a comparison of display strategies on Olivetti(top), Swedish(middle) and Corel(bottom) databases.
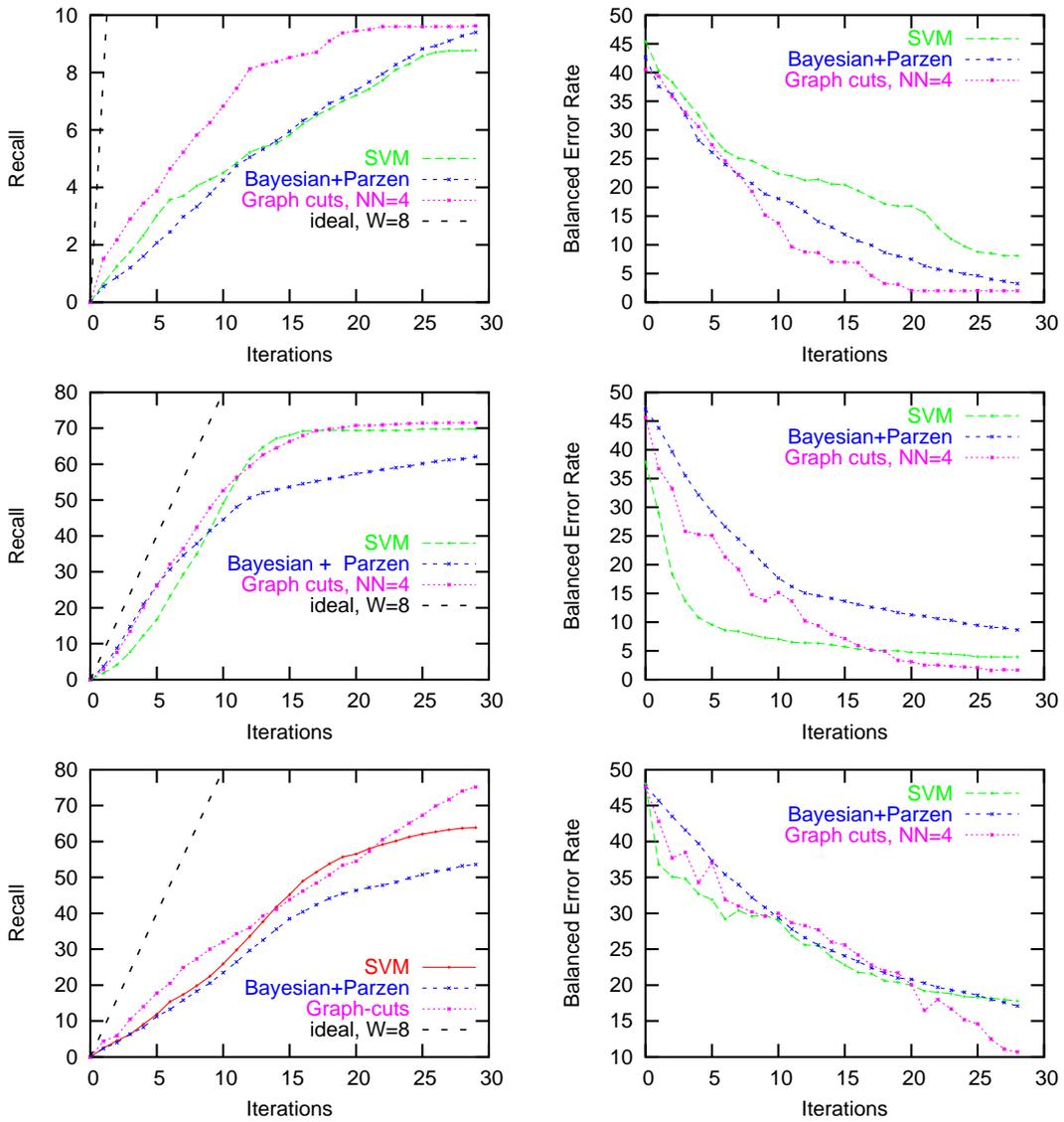
FIG. 7 – This picture shows a comparison of graph-cuts with respect to SVMs and Parzen estimator on the Olivetti(top), the Swedish (middle) and the Corel (bottom) databases.

one $\mathcal{T}_t^- = \{(X_i, Y_i) : Y_i = -1\}$ [33]. This hyperplane corresponds to a non-linear decision function :

$$\sum_{i=1}^{|\mathcal{T}_t|} \alpha_i \, Y_i \, k(\|X - X_i\|) \; + \; b \qquad (10)$$

whose sign decides about the membership of a given $X \in \mathcal{X}$. In the above function $\{\alpha_i\}$ are the training parameters and $b \in \mathbb{R}$. Each parameter is chosen such that $\alpha_i \in [0, C]$, where C is the regularization parameter making the trade-off between fitness and complexity of the resulting SVM classifier. In practice, we set C to a very large value (see [11]).

The results reported in Figure (7), show that in almost all the cases, the recall and the generalization performances of relevance feedback (using graph-cuts) are better than SVMs and Parzen based RF. The diagrams in the middle of this figure, shows an exception where SVMs achieved the best performance in generalization but not in recall so again these two measures are not strictly correlated. The outperformance of relevance feedback using graph-cuts appears in (1) the fact that the generalization error reaches its smallest value *mainly* at the end of the interaction process and (2) the fact that the area under the curve is clearly the largest for recall[4]

5 Discussion and Conclusion

We introduced in this work an original approach for relevance feedback based on graph-cuts transducers. This work demonstrates clearly that our variational approach is effective in order to handle transductive learning. The latter shows a clear and consistent improvement with respect to the most powerful and used techniques in relevance feedback including SVMs and Parzen windows. This is achieved by incorporating the structure of the unlabeled data in the training process. In contrast to inductive learning, the propagation of the labels on unseen data is carried out as a part of the training process, thereby improving the generalization and the recall performances.

The success of our relevance feedback system comes also from the display strategy which guides the search process. This search is implemented using a graph which models the topology of the manifold enclosing the training database and helps us to find and explore ambiguous data efficiently. We conclude that for simple databases, containing homogeneous classes, we need only to explore samples close to the decision boundary and this is translated into picking

---

[4]Even though the area under the curve is not computed in this paper, it is easy to spot the difference of this area mainly for recall.
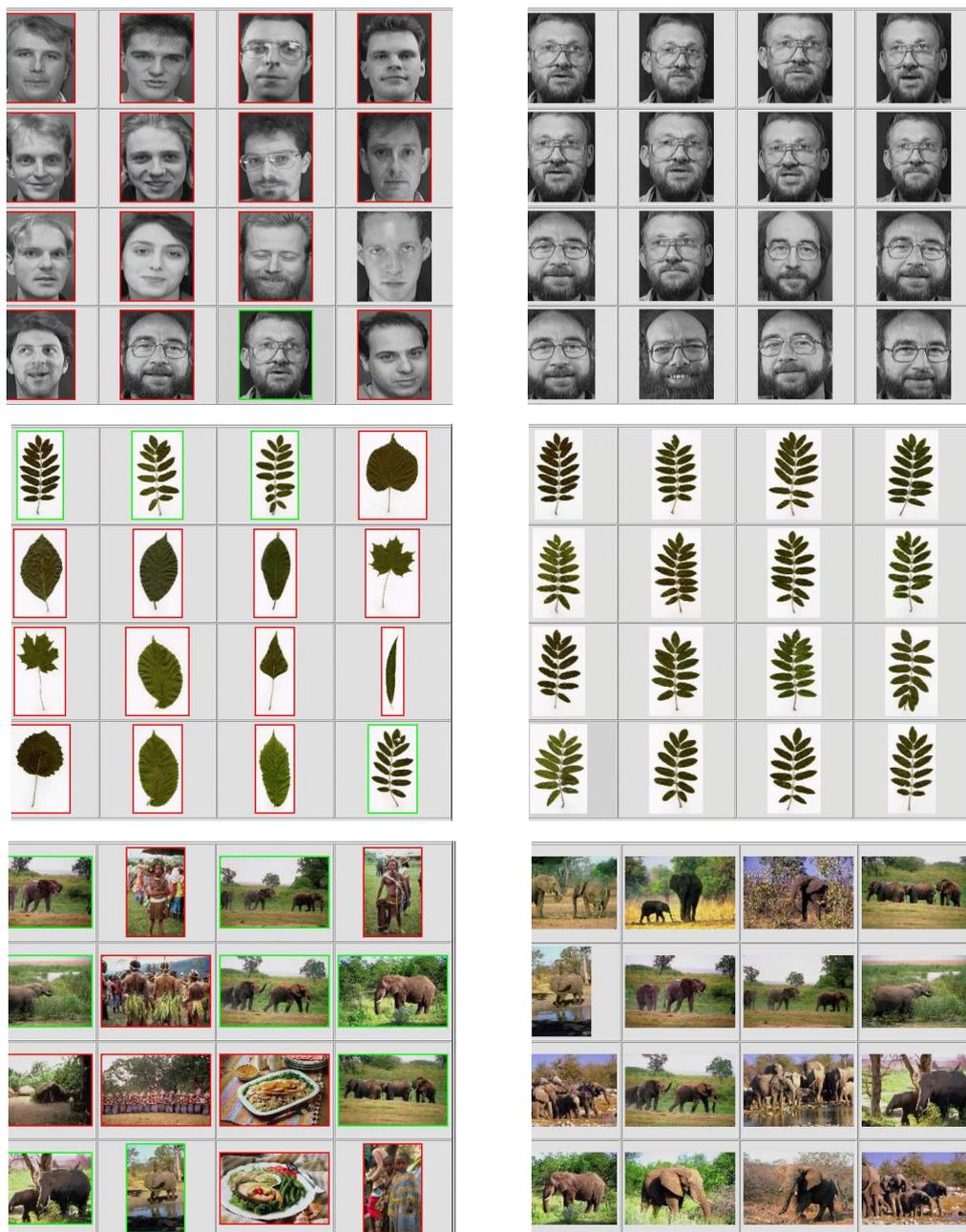
FIG. 8 – An example of an RF session on Olivetti (top) Swedish (middle) and Corel databases (bottom). First results found after submitting the top left image as a query (left) the result of RF after 2, 2 and 5 iterations respectively for Olivetti, Swedish and Corel (right).

the data of the min-cut edges. When showing those images, the user will actively correct the labeling of unseen data and helps the RF system to refine the decision boundary locally. On the other-hand when the database is difficult, i.e., contains spread classes with several modes it is important to get an active feedback from the user on both ambiguous images as well as on a subset taken *randomly*, i.e. both exploitation and exploration are then required.

Other different settings were explored in this paper including the choice of the topology (choice of the degree of the graph) and the kernel similarity function. Indeed, the choice of the kernel is very prominent and the triangular kernel is a good candidate. Beside its rotation and translation invariance, also achieved by the Gaussian one, our proposed kernel is also scale invariant and in practice it shows good performance in RF. Notice that we limited our comparison (to SVMs and Parzen windows) using only this kernel since previous works have shown that the triangular kernel outperforms significantly the traditional Gaussian kernel.

Finally, as a future work, we will investigate the application of this method to large scale databases and incremental graph-cuts for efficient training and classi-fication.

Références

[1] F.R. Bach. Active learning for misspecified generalized linear models. *Advances in Neural Information Processing Systems (NIPS)*, 19, 2006.

[2] N. Boujemaa, F. Fleuret, V. Gouet, and H. Sahbi. Visual content extraction for automatic semantic annotation of video news. *In the proceedings of the SPIE Conference, San Jose, CA*, 2004.

[3] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *In IEEE transactions on Pattern Analysis and Machine Intelligence*, 23(11) :1222–1239, 2001.

[4] G. Caenen and E.J. Pauwels. Logistic regression model for relevance feedback in content based image retrieval. *In proceedings of SPIE*, 4676 :49–58, 2002.

[5] Y. Chen, X.S. Zhou, and T.S. Huang. One-class svm for learning in image retrieval. *Int'l Conf on Image Processing*, 2001.

[6] I.J. Cox, M. Miller, T.P. Minka, T. Papathomas, and P. Yianilos. An optimized interaction strategy for bayesian relevance feedback. *IEEE Conf. Computer Vision and Pattern Recognition*, 1998.

[7] I.J. Cox, M. Miller, T.P. Minka, T. Papathomas, and P. Yianilos. The bayesian image retrieval system, pichunter : theory, implementation, and psychophysical experiments. *IEEE Trans. On Image Processing*, 9(1) :20–37, 2000.

[8] I.J. Cox, M.L. Miller, T.P. Minka, and P.N. Yianilos. An optimized interaction strategy for bayesian relevance feedback. *IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, California*, pages 553–558, 1998.

[9] Y. Fang and D. Geman. Experiments in mental face retrieval. *Proceedings AVBPA 2005, Lecture Notes in Computer Science*, pages 637–646, July 2005.

[10] M. Ferecatu, M. Crucianu, and N. Boujemaa. Retrieval of difficult image classes using svd-based relevance feedback. *MIR '04 : Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 23–30, 2004.

[11] F. Fleuret and H. Sahbi. Scale-invariance of support vector machines based on the triangular kernel. *Third International Workshop on Statistical and Computational Theories of Vision (part of ICCV2003) , Nice, October*, 2003.

[12] L.R. Ford and D.R. Fulkerson. Flows in networks. *Princeton University Press*, 1962.

[13] D. Geman and R. Moquet. A stochastic model for image retrieval. *In Proceedings of RFIA*, 2000.

[14] Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader : query databases through multiple examples. *Int'l Conf. on Very Large Data Bases (VLDB), NY*, 1998.

[15] Olivier Juan and Yuri Boykov. Active graph cuts. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 1023–1029, 2006.

[16] K.I. Kim, K. Jung, and H.J. Kim. Face recognition using kernel principal component analysis. *Signal Processing Letters*, 9(2) :40–42, 2002.

[17] Pushmeet Kohli and Philip H. S. Torr. Measuring uncertainty in graph cut solutions - efficiently computing min-marginal energies using dynamic graph cuts. In *ECCV*, pages 30–43, 2006.

[18] V. Kolomogorov and R. Zabih. What energy functions can be minimized via graph cuts. *in IEEE transactions on Pattern Analysis and Machine Intelligence*, 26(2) :147–159, 2004.

[19] T. Kurita and T. Kato. Learning of personal visual impression for image database systems. *In the proceedings of the international conference on Document Analysis and Recognition*, 1993.

[20] S.D. MacArthur, C.E. Brodley, and C. Shyu. Relevance feedback decision trees in content-based image retrieval. *IEEE Workshop CBAIVL*, 2000.

[21] C. Meilhac, M. Mitschke, and C. Nastar. Relevance feedback in surfimage. *Proceedings of the 4th IEEE Workshop on Applications of Computer Vision*, 1998.

[22] H. Narayanan and M. Belkin. On the relation between low density separation, spectral clustering and graph cuts. *Advances on Neural information processing systems NIPS*, 2006.

[23] R.W. Picard, T.P. Minka, and M. Szummer. Modeling user subjectivity in image libraries. *In the proceedings of the international conference on Image Processing*, 1996.

[24] F.R. Preparata and M.I. Shamos. Computational geometry : An introduction. *New York : Springer-Verlag*, 1985.

[25] Y. Rui, T.S. Huang, and S. Mehrotra. Relevance feedback techniques in interactive content-based image retrieval. *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 25–36, 1998.

[26] Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback : A power tool in interactive content-based image retrieval. *IEEE Tranactions on Circuits and Systems for Video Tech*, 8(5) :644–655, 1998.

[27] H. Sahbi. Kernel pca for similarity invariant shape recognition. *In the Journal of Neurocomputing*, 2006.

[28] G. Schohn and D. Cohn. Less is more : Active learning with support vector machines. *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 839–846, 2000.

[29] B. Scholkopf, R. Williamson, A. Smola, J.S. Taylor, and J. Platt. Support vector method for novelty detection. *Adv. in Neural Information Processing Systems, MIT Press.*, 2000.

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

[31] K. Tieu and P. Viola. Boosting image retrieval. *IEEE Conf. Computer Vision and Pattern Recognition*, 2000.

[32] S. Tong and E. Chang. Support vector machine active learning for image retrieval. *MULTIMEDIA '01 : Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118, 2001.

[33] Vladimir N. Vapnik. Statistical learning theory. *A Wiley-Interscience Publication*, 1998.

[34] N. Vasconcelos and A. Lippman. Learning from user feedback in image retrieval. *in Neural Information Processing Systems MIT press*, 2000.

[35] M. Worring, A. Smeulders, and S. Santini. Interaction in content-based image retrieval : a state-of-the-art review. *Int'l Conf. on Visual Info. Sys.*, 2000.

[36] Y. Zhao, Yao Zhao, and Z. Zhu. Relevance feedback based on query refining and feature database updating in cbir system. *Signal Processing, Pattern Recognition, and Applications*, 2006.

[37] X.S. Zhou and T.S. Huang. Relevance feedback in image retrieval : A comprehensive review. *in IEEE CVPR Workshop on Content-based Access of Image and Video Libraries (CBAIVL)*, 2006.