Multimedia Systems (2009) 15:201–219
DOI 10.1007/s00530-009-0154-4

REGULAR PAPER

Visual and textual fusion for semantically supervised region-based retrieval

Rongrong Ji · Hongxun Yao · Pengfei Xu · Xiaoshuai Sun

Received: 13 January 2008 / Accepted: 6 May 2009 / Published online: 20 June 2009

Abstract This paper presents a unified annotation and retrieval framework, which integrates region annotation with image retrieval for performance reinforcement. To integrate semantic annotation with region-based image retrieval, visual and textual fusion is proposed for both soft matching and Bayesian probabilistic formulations. To address sample insufficiency and sample asymmetry in the annotation classifier training phase, we present a region-level multi-label image annotation scheme based on pair-wise coupling support vector machine (SVM) learning. In the retrieval phase, to achieve semantic-level region matching we present a novel retrieval scheme which differs from former work: the query example uploaded by users is automatically annotated online, and the user can judge its annotation quality. Based on the user's judgment, two novel schemes are deployed for semantic retrieval: (1) if the user judges the photo to be well annotated, *Semantically supervised Integrated Region Matching* is adopted, which is a keyword-integrated soft region matching method; (2) If the user judges the photo to be poorly annotated, *Keyword Integrated Bayesian Reasoning* is adopted, which is a natural integration of a *Visual Dictionary* in online content-based search. In the relevance feedback phase, we conduct both visual and textual learning to capture the user's retrieval target. Better annotation and retrieval performance than current methods were reported on both *COREL 10,000* and Flickr web image database (25,000 images), which demonstrated the effectiveness of our proposed framework.

Keywords Image annotation · Image retrieval · Relevance feedback · Visual and textual fusion · Multi-label annotation · Pairwise coupling · Bagging · Bayesian formulation · Soft region matching · Visual dictionary · Visual and textual learning

1 Introduction

With the prevalence of digital cameras, rapid advances in web technology have facilitated the generation and storage of image collections. Effective retrieval of images from these gigantic collections poses significant technical challenges. Content-based image retrieval (CBIR) aims to retrieve images based on their visual contents. This has been an active research area over the past decade [18,22]. However, the performance of current CBIR systems is still unsatisfactory because of the semantic gap between the visual contents and the human perception of an image. Many efforts have been made to bridge this semantic gap, through either online interaction [i.e. *relevance feedback* (RF)] or offline training (i.e. *image annotation)*. Although progress has been made, the state of the art is still unsatisfactory for real-world application, because of difficulties in understanding image contents and the retrieval intentions of the user.

One feasible solution is to annotate the visual contents of images with keywords. Although the volume of image collections restricts their coverage by manual keyword labeling, computerized automatic image annotation [2,7,11,13–

Communicated by Mohan Kankanhalli, Ph.D.

R. Ji (✉) · H. Yao · P. Xu · X. Sun
Department of Computer Science, Harbin Institute of Technology,
No. 92, West Dazhi Street, P.O. BOX 321, 150001 Harbin,
People's Republic of China
e-mail: rrji@vilab.hit.edu.cn

H. Yao
e-mail: yhx@vilab.hit.edu.cn

P. Xu
e-mail: pfxu@vilab.hit.edu.cn

X. Sun
e-mail: xssun@vilab.hit.edu.cn

16] is a promising solution to facilitate image search. In such a scenario, a manually pre-labeled image collection is provided beforehand to train a model that captures the semantic associations between annotation keywords and image visual contents. These associations are used to propagate keywords to the remaining unlabeled images in the database. For instance, Chang et al. [2] developed the CBSA system with Bayesian Point Machine for automatic image annotation. Goh et al. [8] used a binary classifier combination strategy, in which a confidence-based dynamic ensemble scheme was adopted to adjust annotation confidence. Lu et al. [16] adopted a heuristic semantic network to associate keywords with images in RF learning. Similar work by Jing et al. [13] adopted a keyword network for dual-interface [both query-by-keyword (QBK) and query-by-example (QBE)] image retrieval.

Another feasible solution is to index and retrieve image content at the object level. Region-based image representation is frequently used to achieve this goal [1,3,15,23,24]. For instance, Carson et al. [1] adopted region-based query to precisely partition a query example into target query regions that better represented the user target in an image search. Wang et al. [23] proposed an integrated region matching (IRM) algorithm. By smoothing over the imprecise distance representation, IRM used soft region matching to ensure robustness against inaccurate segmentation. Wang et al. [24] proposed a constraint-based region matching approach to integrate spatial information into IRM. However, the semantic gap still exists due to the difficulties in image segmentation and semantic perception.

Recently, the advantages of both solutions were combined, with the aim of keyword annotation at both the object and region levels to precisely capture image semantics. For modeling the statistics of regions and words, the associations between keywords and image contents (especially region contents) are learned by statistical textual-visual correlation. Representative work includes the translation model (TM) [7], cross-media relevance model (CMRM) [11], continuous relevance model (CRM) [14], and two-dimensional multi-resolution hidden Markov model (2D-MHMM) [15]. For a detailed review, please refer to the survey of current progress in CBIR in [6].

Two major drawbacks exist in these approaches: first, their methods try to propagate annotations at the image level rather than the object level, which is unsuitable when the same objects are surrounded by diverse backgrounds. Second, how to use the annotation results to improve retrieval has not been adequately explored.

In this paper, we propose to seamlessly fuse image annotations with the image retrieval process in order to bridge the semantic gap in CBIR. To the best of our knowledge, the combination of text annotations with image contents for retrieval is still not well explored. Simple linear combination is prevalent in previous work [13,16]. Datta et al. [5] attempted to integrate annotation results with image retrieval, retrieval was driven by WordNet-based bag-of-words distances of image tags, and annotation was performed using image categorization. Two scenarios were presented in their work: (1) annotate the whole database and then conduct WordNet-based keyword search, (2) expand the query keyword into query examples and conduct content-based search. However, the fusion of visual and textual information into a unified ranking framework was not investigated. We believe that seamless visual and textual fusion strategy is needed to effectively integrate annotations into CBIR for real-world application.

We present a unified region annotation and retrieval framework to achieve this goal. For offline annotation of database images, we present a pairwise coupling (PWC) support vector machine (SVM) bagging network to annotate the training keywords over the image database at the region level, in which each region is annotated with multiple labels with confidence weights. In online retrieval, we present a novel user-computer interaction scheme: the annotation-interaction-retrieval process for CBIR. It uses automatic image annotation as a pre-step, based on which the user can judge the quality of annotation results (well annotated or poorly annotated). Using user preference, two fusion strategies are deployed to combine textual and visual information for retrieval. If the query image is well-annotated, we present a keyword-integrated soft region matching strategy, named *semantically supervised integrated region matching* (SSIRM). SSIRM combines the advantages of soft matching and linguistic supervision. If the query image is poorly annotated, we present a statistical textual-visual correlation analysis strategy named *keyword-integrated Bayesian reasoning* (KIBR). KIBR exploits the fusion strategy in a probabilistic Bayes formulation, which uses the semantic information to "bridge" the semantic gap. Finally, in RF, visual and textual learning is achieved to capture users' retrieval targets at a higher semantic level.

Figure 1 presents the system flowchart of our proposed framework. In the image annotation process, to address the sample asymmetry and sample insufficiency issues in keyword classifier learning, inter-keyword (inter-class) SVM classifiers are trained by bagging training samples. Multi-label annotation is achieved using the trained PWC SVM bagging network for keyword propagation. We further present a novel annotation-interaction-retrieval CBIR process for visual and textual fusion; when a user makes a query, a pre-annotation step is introduced to capture higher level user semantics. Based on the labeling judgment, semantic-level retrieval is achieved by integration of the users' context and content supervision. Depending on the annotation effectiveness, multi-labeled region keywords are integrated into two region retrieval schemes: (1) the "semantic factors" scheme to affect the region-based soft matching in SSIRM

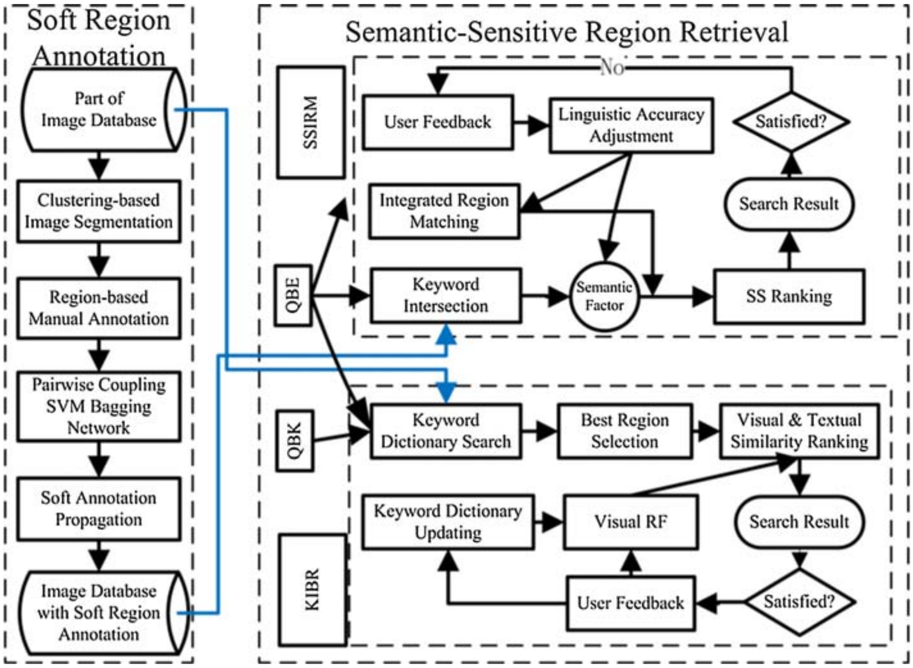

Soft Region Annotation
Part of Image Database
Clustering-based Image Segmentation
Region-based Manual Annotation
Pairwise Coupling SVM Bagging Network
Soft Annotation Propagation
Image Database with Soft Region Annotation
Semantic-Sensitive Region Retrieval
SSIRM
QBE
QBK
KIBR
No
User Feedback
Linguistic Accuracy Adjustment
Satisfied?
Integrated Region Matching
Search Result
Keyword Intersection
Semantic Factor
SS Ranking
Keyword Dictionary Search
Best Region Selection
Visual & Textual Similarity Ranking
Keyword Dictionary Updating
Visual RF
Search Result
User Feedback
Satisfied?

Fig. 1 Visual and textual fusion framework for region annotation and retrieval

and (2) the "visual dictionary" scheme to associate image visual contents in KIBR using a probability-based Bayes formulation. In the RF process, we optimize both content-level RF learning and context-level RF learning and fuse their predictions, which provides our system with more precise query concept description at the semantic level.

Comparing with the state-of-the-arts, there are three contributions in our paper:

1. From the algorithmic viewpoint, we explore the fusion of visual and textual information from both soft reasoning and Bayesian reasoning for performance reinforcement, which is of fundamental importance for improving current CBIR systems. In the annotation phase, we adopt the PWC-SVM bagging network to cope with the problems of sample insufficiency and sample asymmetry in learning. In the retrieval phase, SSIRM can largely reduce the negative effect of imprecise segmentation by soft matching; KIBR supports both QBK and QBE interfaces.
2. From the perspective of interactivity, we present a novel annotation-interaction-retrieval CBIR process, which is different from all current user-computer interactions in CBIR. This novel process allows us to precisely capture the user's target for retrieval and to deploy appropriate schemes (SSIRM or KIBR) for different qualities of annotation so that the system can interpret the query example.
3. For experimental validation, both a comparison with the state-of-the-arts in a standard database (*COREL 10,000*) and a real-world application in a Flickr web image database are evaluated in this paper, demonstrating our advantages in quantized and real-world scenarios.

2 Image segmentation and training collection construction

We adopted the clustering-based image segmentation algorithm used in SIMPLIcity [23], which has been shown to be effective and efficient for real-time systems. Since we need online query segmentation as well as offline database processing, efficiency is important. The clustering procedure automatically clusters grid-based image blocks into regions. An image is divided into grid blocks. The size of each block is 4×4 pixels, which was the empirical setting in [23] and also our empirical choice in our experimental scenario. In each segmented region a six-dimensional feature is extracted as follows:

The first three dimensions are the averages of the H, S and V components, respectively, in HSV color space; the other three are the variances of the HH, HL, and LH high-frequency coefficients of Daubechies four wavelet bands from the H component in each block, similar to [23].

Based on this six-dimensional feature vector, the k-mean clustering process is conducted to group blocks into regions (equivalently, segment the image into regions). This procedure does not specify the cluster number, but gradually iterates this number from 2 to 6, seeking the best image partition by maximizing this evaluation criterion:

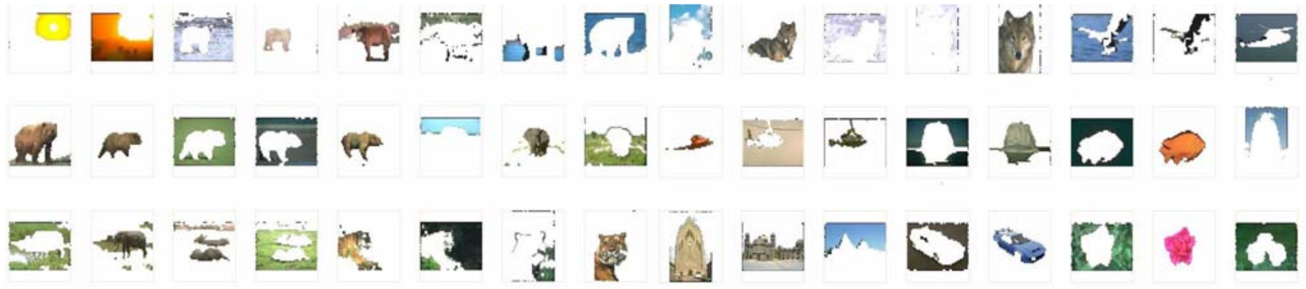


Fig. 2 Image segmentation based on pixel clustering

$$N_{\text{Selected}} = \arg\min_i \left(\frac{1}{N_i} \sum_{j=1}^{N_i} \left\| F_j^{\text{Query}} - F_j^{\text{Image}} \right\|_{L2} \right) \quad (1)$$

where F_j^{Query} is the jth six-dimensional feature vector of the query image, F_j^{Image} is the jth six-dimensional feature vector of the image in the current segmentation, and N_i is the ith iteration (2–6). The rationale for Eq. 1 is that the segmentation results should be as similar as possible to the query. In online query, the query image is pre-segmented into four regions. The 2–6 segmentations of the overall database are pre-processed and stored offline to improve online efficiency. This region number is our empirical compromise since users are usually concerned with less than three foreground objects in our experiments.

The dissimilarity between two regions is defined as their $L2$ distance in the feature space. Different from the features used for segmentation, the features used for region description can contain information more meaningful in perception. To describe the region more effectively, our algorithm also uses shape features, as was done in [23]: the geometrical circularity (Eq. 2) and its center scatter of this region (Eq. 3). This addition extends our feature representation of regions to eight dimensions.

$$\alpha = \frac{4\pi S}{P^2} \quad (2)$$

$$\mu_{p,q} = \frac{1}{n_{\text{Block}}} \sum_{(x,y)\in R} (x - x_c)^2 (y - y_c)^2 \quad (3)$$

In Eq. 2, α represents the geometrical circularity of this region; S and P are its size and perimeter respectively. In Eq. 3 $\mu_{p,q}$ is the center scatter of this region, and x_c and y_c and n_{Block} represent the geometric centers and the number of blocks in region R, respectively. As presented in Fig. 2, the segmentation results are visually similar and can be spatially unconnected. This is useful to separately describe background or occluded objects.

Subsequently, each region in this segmented image collection is manually labeled with a single keyword; their ensemble forms our *visual dictionary* as labeling ground truth. The semantic representation ability of the visual dictionary is co-determined by both the number of keywords and the richness of the segmented image collection. In our experiment, 2,000 segmented regions were collected from the COREL database for manual labeling, and 202 keywords were selected to construct the visual dictionary. It is clear that this visual dictionary still cannot successfully describe the semantic information in each image in a real-world application. Two problems would arise, namely, sample insufficiency in the annotation classifier training and keyword insufficiency in user interactions. These two problems will be further addressed in the following sections.

3 Multi-label region annotation based on pairwise coupling SVM bagging network

A region-level, classification-based annotation strategy is adopted in this paper. We adopt the SVM to construct our keyword annotation classifier, as is common in CBIR [19,20]. It aims to separate two classes of training samples in the feature space by an optimal hyperplane, where the maximum geometric margin is expected. However, an SVM classifier is unstable when the training samples are insufficient, and overfitting would occur [19]. Furthermore, its optimal samples may be biased when the volumes of the positive and negative samples are asymmetric. In image annotation, the manually labeled regions are often insufficient for SVM training. Moreover, it cannot be guaranteed that the volumes of training samples for two keywords are strictly symmetric. Therefore, both sample insufficiency and sample asymmetry problems exist in SVM training. This would result in low accuracy in subsequent annotation propagation.

To solve these two problems, we present a bagging process that constructs an integrated classifier for inter-keyword (inter-class) SVM classification, which is achieved by training sample bootstrapping and SVM aggregation. Here we explain the bagging-based construction of the training set:

(1) When the numbers of positive and negative examples are small, training several SVM classifiers by bagging

Table 1 Inter-class keyword classifier bagging

```
Input: training samples of keyword $K_i$ and $K_j$ : $S_i$ and $S_j$, weak classifier $C$ (SVM), and bagging iteration $T_1$
 For bootstrapping interval $k$ from 1 to $T_1$
 Begin
     Bootstrap two partial training sample collections: $S^+$ from $S_i$ and $S^-$ from $S_j$, subject to $|S^+| = |S^-|$.
          Train $k$th SVM $C_k$ based on positive and negative samples
 End
 Construct $C^{ij}$ by aggregation of all $C_k$
Output: $C_{ij}$ to classify keywords $i$ and $j$
```

Table 2 PWC SVM classification for annotation

```
Input: test regions $R^t$ for annotation, pair-wise SVM classifiers $C_{01}, C_{02}, \ldots, C_{m(m-1)}$, weak classifier $C$ (SVM)
  For each keyword $K_i$
      Use its $(m-1)$ corresponding SVM classifiers $C_{i1}, C_{i2}, \ldots, C_{i(m-1)}$ to classify the test region's membership to this keyword
      Calculate its pairwise coupling result $V_i$ by this $(m-1)$ SVM voting
Output: Use the $n$ keywords with the $n$ highest classifier outputs to annotate this region, with the corresponding voting result $V$ as the
confidence factors of these annotated keywords
```

training examples would generate more stable results than one SVM. As a discriminative learning scheme, one SVM is more likely to be affected and biased by data outliers. With several SVM classifiers, one prediction error would not cause too serious effects since the SVM classifiers are combined in a voting committee. In other words, overfitting would be less likely to happen with multiple SVMs.

(2) As pointed out by Tao et al. [19] in their PAMI 06 paper, when the positive and negative training examples are unbalanced, the SVM would shrink toward the label with fewer training examples. To address this issue, we adopted a scheme that is similar to [19], in which we conduct bagging over the training examples with larger number (e.g. negative examples), to produce several equal-sized training sets (positive training number = negative training number). Consequently, the learning results of each SVM classifier are more stable and closer to the "real" classifier hyperplane. Finally, the outputs of these SVM classifiers are combined to produce the final results.

This bagging process is presented in Table 1.

Furthermore, as a binary classifier, the SVM is unsuitable for direct application in annotation propagation, which is a multi-class classification problem. Generally speaking, there are two strategies to extend binary classifiers to a multiple-class situation:

1. One-to-all: For each class, train an intra-class classifier whose classification result indicates the confidence that the test example belongs to this class (between [0, 1]). For each test sample, these classifiers of all categories are used to produce the classification results, and the class with the highest output is selected as the classification result.
2. One-to-one: For each category, train $(m-1)$ pairwise classifiers (inter-class) between the current class and the other classes (totally m classes). For each test sample, the confidence of each class is the majority voting result of its corresponding $(m-1)$ classifiers (each between [0, 1]). The class with the highest accumulated output is selected as the classification result. This is called PWC, which combines the outputs of all classifiers for prediction. It has gained more acceptance due to its good generalization ability [17,25].

We adopted the PWC strategy to train multi-class SVMs for annotation classification. We train $(m-1)$ inter-keyword SVMs for each keyword to determine its annotation confidence of the test region (totally C_m^2 SVM classifiers). The top n keywords with the highest SVM voting results are selected to annotate this region. The PWC SVM classification algorithm is presented in Table 2.

4 Visual and textual fusion for region retrieval

We adopt annotation propagation results as semantic supervision to enhance retrieval performance. In previous work [16,13], the integration of text annotation and image contents was not well explored, in which simple linear combination of visual and textual similarity was prevalent. This paper

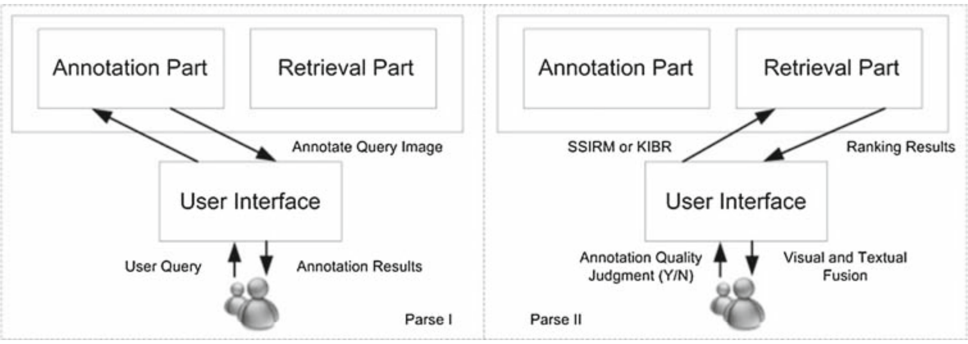


Fig. 3 Typical retrieval scenario of the proposed visual and textual fusion framework

Table 3 Semantic-supervised integrated region matching

For each region i in the query image and each region j in the test image, **evaluate** the "semantic factor" w_{ij} by Eq. 6

Set the assigned region set *L={}*. **Set** the unassigned region set $E = \{(i, j)|i = 1, 2, \ldots, M; j = 1, 2, \ldots, N\}$

Calculate the *weighted distance* between regions i and j, in which weight w_{ij} is calculated by Eq. 6

Choose the minimum d_{ij} for $(i, j) \in E - L$. **Label** the corresponding (i, j) as (i', j')

$\min(p_{i'}, p'_{j'}) \rightarrow w_{i'j'}$

If $p_{i'} < p^{'}_{j'}$, **set** $w_{i'j} = 0, j \neq j'$; otherwise **set** $w_{ij} = 0, i \neq i'$

$p_{i'} - \min(p_{i'}, p'_{j'}) \rightarrow p_{i'}$

$p'_{j'} - \min(p_{i'}, p'_{j'}) \rightarrow p'_{j'}$

$L + \{(i', j')\} \rightarrow L$

If $\sum_{i=1}^{M} p_i > 0 \&\& \sum_{j=1}^{N} p'_{j'} > 0$, **go to** step 2; otherwise **stop**

explores the fusion of visual and textual information to reinforce each other in similarity ranking. The key idea of our strategy is a three-phase annotation-interaction-retrieval model. Figure 3 presents a typical retrieval scenario in our system:

Generally speaking, our search process can be described as follows: (1) the user uploads a query example (photo) to our system, (2) the system annotates the query example online and presents the annotation results to the user interface, (3) the user judges the annotation results as either well annotated (satisfied) or poorly annotated (unsatisfied), then selects the "*well-annotated*" or "*poorly-annotated*" check box in our interface, and (4) the system selects the search schemes based on the user preference, in which the "*well-annotated*" choice corresponds to SSIRM and "*poorly-annotated*" corresponds to KIBR.

In the case that the query image is well-annotated, a straightforward strategy is to adjust the region-level similarity by a "semantic factor". Based on this idea, we extended IRM [23] by linguistic association to SSIRM.

In the case that the query image is poorly annotated, we treat the ground-truth manual annotation set as a *visual dictionary* to "translate" the image representation into a keyword representation and then "reverse-translate" it back again for similarity ranking. We present a probabilistic Bayesian reasoning framework: KIBR to "bridge" the semantic gap by inference through keywords.

4.1 Semantic-supervised integrated region matching (SSIRM)

Once a query image is provided by users, it is first segmented into regions. For region-based similarity calculation, our scheme extends the IRM [23] algorithm by adding semantic supervision (SSIRM). The IRM uses the "most similar highest priority" (MSHP) principle to estimate W_{ij} with Eqs. 4 and 5, which can effectively mitigate the negative influence of imprecise image segmentation.

However, in the IRM, no semantic supervision is used to reveal the user's image perception. Our improvement of IRM can be described in Table 3, in which the following constraints hold (p_i and p'_j demonstrate the normalized importance of regions r_i and r'_j in each image, respectively):

$$\text{Similarity}(I_1, I_2) = \sum_{i=1}^{M} \sum_{j=1}^{N} w_{ij} S(r_i, r\prime_j)$$

$$\text{s.t.} \quad \sum_{i=1}^{M} \sum_{j=1}^{N} w_{ij} = 1 \tag{4}$$

$$\sum_{i=1}^{M} p_i = \sum_{j=1}^{N} p'_j = 1 \quad \sum_{i=1}^{M} w_{ij} = p'_j, \; \sum_{j=1}^{N} w_{ij} = p_i \tag{5}$$

Most part of SSIRM is identical to IRM except that the initial weight between two regions "w_{ij}" is calculated based on the similarity of corresponding semantic annotation results as in Eq. 6. In SSIRM (Table 3), the region annotation results are used to supervise the region matching process. Specifically, a keyword labeling intersection procedure is proposed to define the "semantic factor" of the semantic similarity between two matching regions (Eq. 6):

$$w_{ij} = \sum_{k=1}^{m} \min(V_{ik}, V_{jk}) \tag{6}$$

where w_{ij} is the weight of two regions between two images in Eq. 4; and V_{ik} and V_{jk} are the confidence factors of the kth keyword (voting results) calculated using the PWC algorithm in Table 3. Then we normalize all w_{ij} for similarity calculations. This weight (semantic factor) reflects the integration of the semantic annotation results in region similarity matching based on Eq. 6. Unlike the calculation of w_{ij} in IRM, which does not involve any semantic-level similarity considerations, our algorithm takes the semantic similarity as a core consideration in region-based similarity matching.

As an intuitive interpretation, the calculation of the semantic factor w_{ij} is similar to the histogram intersection. If two regions have similar keyword annotations with high keyword confidence factors, their similarity would affect the image similarity more.

Using this initial region weighting, the IRM process is conducted to calculate the region-level similarity between two images. This semantic-supervised region matching process is conducted between the query image and all the images in the image database. The top t images with the t highest similarities are returned to users as the retrieval result of the current interaction iteration.

This visual-level region matching process is identical to the IRM algorithm, except that the initial weights w_{ij} of corresponding regions i and j is calculated by the semantic factor evaluated in Eq. 6.

4.2 Keyword-integrated Bayesian reasoning (KIBR)

Bayesian reasoning has been investigated by researchers [4], [12] to facilitate CBIR. For instance, the PicHunter system [4] used a Bayesian framework to model users' action and then predicted the retrieval results of target images. Zhang et al. [27] proposed a probabilistic hidden semantic model for region-based semantic representation and used its posterior probability for image retrieval. However, in previous work the Bayesian method has been exploited to infer only the visual correlations between global image characteristics and text annotations.

In our previous work [12], region-based image features were combined with keyword-based image perception to add semantic information into the similarity calculation and ranking process. However, the reasoning phase of [12] was confined to the initial similarity ranking, and no semantic supervision was considered in RF. By fusing visual and textual information to improve similarity ranking in this paper, we not only improve our former reasoning framework, but also introduce the Bayesian reasoning integration of KIBR into RF learning. In this subsection, we present our Bayesian reasoning idea, which supports both QBE and QBK scenarios. For these two query patterns, we present our Bayesian reasoning process in detail below:

4.2.1 Bayesian reasoning for query-by-example scenarios

From a probabilistic conditional inference viewpoint, *QBE* in CBIR can be regarded as a process to deduce the conditional probability $P(I|I_Q)$ of each image I in the database given query example I_Q. Given the query example I_Q, this probability measures to what degree an image I similar to query image I_Q. In other words, given query image I_Q, $P(I|I_Q)$ reveals the probability that a retrieval system will find image I based on their content and semantic similarity. Subsequently, given the query example, the similarity ranking process is conducted by ranking the conditional probability of each image. To measure this probability, an equivalent transformation is made by replacing the query image I_Q by its regions: $R_1, R_2, \ldots, R_{\text{query}}$ as Eq. 7:

$$P(I|I_Q) = P(I|R_1, R_2, \ldots, R_{\text{query}}) \tag{7}$$

Using the visual dictionary, the Bayes posterior probability of semantic similarity (for each region in the image database given the query image) can be expanded into Eq. 8 as a discrete integral over variable K using Bayesian reasoning:

$$P(I|I_Q) = \sum_{K=1}^{K_{\text{dic}}} P(I, K|R_1, R_2, \ldots, R_{\text{query}}) \tag{8}$$

in which K means the Kth keyword from our annotation dictionary, and K_{dic} is the number of keywords in this dictionary. Based on conditional independence between I and K, Eq. 8 can be expanded into:

$$\begin{aligned} P(I|I_Q) = \sum_{K=1}^{K_{\text{dic}}} & P(I|K, R_1, R_2, \ldots, R_{\text{query}}) \\ & \times P(K|R_1, R_2, \ldots, R_{\text{query}}) \end{aligned} \tag{9}$$

The second component of Eq. 9 is further transformed by another Bayes formula procedure into Eq. 10:

$$\begin{aligned} & P(K|R_1, R_2, \ldots, R_{\text{query}}) \\ & \quad = \frac{P(R_1, R_2, \ldots, R_{\text{query}}|K)P(K)}{P(R_1, R_2, \ldots, R_{\text{query}})} \end{aligned} \tag{10}$$

We assume that the posterior probabilities of the regions given their keyword annotation are conditionally independent. Using a naïve Bayesian classifier, the probability

$P(R_1, R_2, \ldots, R_{\text{query}}|K)$ can be considered as the product of individual keyword-to-region probabilities:

$$P(R_1, R_2, \ldots, R_{\text{query}}|K) = \prod_{i=1}^{\text{query}} P(R_i|K) \tag{11}$$

For each region in the query image, its eight-dimensional feature vector is extracted using the method described in Sect. 2 (This process is done online and the former six-dimensional feature is used for segmentation). For each keyword, the most similar region annotated with this keyword in the visual dictionary is chosen to calculate the probability $P(R_i|K)$ in Eq. 11 as follows:

$$P(R_i|K) = e^{-\left(\sum_{j=1}^{8}\left(\text{Feature}_j^{R_i} - \text{Feature}_j^{R_Q}\right)^8\right)^{\frac{1}{8}}} \tag{12}$$

in which the exponential value of Euclidean distance between region R_i and R_Q is calculated as the similarity between query region R_Q and region R_i in the database.

Similarly, $P(I|K, R_1, R_2, \ldots, R_{\text{query}})$ is derived in the following equations:

$$\begin{aligned} P(I|K, R_1, R_2, \ldots, R_{\text{query}}) &= P(I|K) \\ &= \frac{P(K|R_1, R_2, \ldots, R_n)P(R_1, R_2, \ldots, R_n)}{P(K)} \end{aligned} \tag{13}$$

$$P(R_1, R_2, \ldots, R_n|K) = \prod_{i=1}^{n} P(R_i|K) \tag{14}$$

In Eq. 13, once a specific keyword is given in the similarity calculation, it is assumed that the probability that image I is similar to the users' retrieval concept given the query image I_Q is no longer related to the query image (that is, the query image is simply a instance to inspire the keyword-based Bayesian reasoning). Hence, an assumption of conditional independence between I and I_Q is made once given a specific K. In Eq. 14 the regions in one image are assumed to be conditionally independent. Consequently the Bayesian optimal classifier we adopted is as Eq. 11.

Finally Eqs. 8–14 are integrated into Eq. 7 to deduce the semantic similarity of image I to query image I_Q. The top 100 images with the highest posterior probabilities are returned to the user as the retrieval result. In summary, our key idea falls into Eqs. 11–13, that leverage the manual annotation results as a visual dictionary to translate semantic similarity into content similarity to produce the final similarity score. In our experiments, we also found that our combination scheme is very effective when the pre-annotation results of the query example are poor. Consequently, we allow user judgment of annotation quality, and in the case that the annotation results are poor, we conduct KIBR to achieve better retrieval results than SSIRM.

4.2.2 *Bayesian reasoning for query-by-keyword scenarios*

Our system also supports query-by-keyword. In the case of query-by-keyword, the Bayesian reasoning process can be described as follows:

$$P(I|K) = P(R_1, R_2, \ldots, R_I|K) = \prod_{I=1}^{I} P(R_i|K) \tag{15}$$

$P(I|K)$ represents the probability that image I is similar to the user's query concept given the query keyword K. It is replaced by its regions in Eq. 15. As assumed above, the regions in an image are conditionally independent, which results in the transformation of $P(R_1, R_2, \ldots, R_I|K)$ to the products of $P(R_i|K)$ using the Bayesian optimal classifier in Eq. 7. $P(R_i|K)$ reflects the degree to which region R_i reveals the retrieval semantic concept given keyword K, and is calculated using the method of Eq. 12. Finally, the top 100 images with the highest posterior probabilities are returned to the user as the retrieval result. The keyword input should be constrained to our annotation keyword set. Otherwise, our system would ask users to provide query images to support *QBE*, which on the other hand would teach our retrieval system the meaning of the current keyword and trigger offline PWC bagging training for annotation propagation.

5 Semantic supervision for relevance feedback learning

We further consider the issue of fusing keyword annotations with RF learning for semantically supervised region retrieval. We consider semantic-level RF learning as a semantic factor evaluation process for combination with visual-level learning to enhance retrieval performance. Corresponding to the annotation combination strategies presented in Sect. 3, two semantic-level learning schemes are exploited in this section, respectively:

5.1 Semantic adjustment in SSIRM relevance feedback

In SSIRM, our RF learning algorithm contains a twofold process to integrate semantic supervision and adjustment with visual-level learning. First, SVM learning is adopted for positive/negative image discrimination at the global level. Second, the semantic factor of each keyword is updated using a region-level annotation evaluation process as in Table 4.

In Eq. 19, S_{Final} represents the final re-ranked similarity; $\text{Label}_{\text{SVM}}^{\text{Pos}} = 1$ if the SVM classifier labels the image as positive, otherwise $\text{Label}_{\text{SVM}}^{\text{Pos}} = 0$.

Furthermore, the modification of "semantic factor" can be applied to both SSIRM and schemes based on global-level features [e.g. feature reweighting (FRE), SVM]. In such cases, the semantic correlation between two global images

Table 4 Relevance feedback learning of region annotation evaluation

```
For each keyword in the keyword set
  Set PosSim = NegSim = 0, representing the positive and negative similarity of this keyword, respectively;
  For each positive feedback image
      For each region in this image
          Select the most similar region in our pre-labeled set with this keyword, and update the PosSim as:

$\mathrm{PosSim} = \mathrm{PosSim} + e^{-||R_{\mathrm{Label}} - R_{\mathrm{Pos}}||_2}$     (16)

  For each negative feedback image
    For each region in this image
        Select the most similar region in our pre-labeled set with this keyword, update NegSim as:

$\mathrm{NegSim} = \mathrm{NegSim} + e^{-||R_{\mathrm{Label}} - R_{\mathrm{Neg}}||_2}$     (17)

  Set the weight $P(K)$ as PosSim/NegSim; Set the similarity ranking method in Eq. 6 as follows:

$w_{ij} = \sum_{k=1}^{m} P(k) \min(V_{ik}, V_{jk}) + 1$     (18)

  Re-calculate the total similarity between query and images in the image database using Table 3
  Combine the ranking results with the SVM classification learning result as in Eq. 19 and re-rank all results:

$S_{\mathrm{Final}} = S_{\mathrm{SSIRM}} / (\mathrm{Label}_{\mathrm{SVM}}^{\mathrm{Pos}} + 1)$     (19)
```

can be considered as the product of the semantic factor "w_{ij}" and the re-ranking result (as in Eq. 20) in which $\mathrm{Similarity}_{\mathrm{Global}}$ means the re-ranked global similarity using a certain global-level learning algorithm:

$$\mathrm{Similarity}(I_1, I_2) = \left(\sum_{i=1}^{M} \sum_{j=1}^{N} w_{ij} \right) * \mathrm{Similarity}_{\mathrm{Global}} \quad (20)$$

Section 6 further investigates the performance of (1) *SSIRM* versus *IRM + global-feature-based SVM*; (2) *semantic factor evaluation + FRE* versus *global-feature-based FRE*. A challenging task in our comparison with methods based on global features is that we compare our *semantic factor evaluation + FRE* with three state-of-the-art SVM-based learning schemes. It has already been demonstrated that SVM outperforms FRE by a large margin in RF learning [19,20]. However, as shown in our experiments, by integrating our semantic learning, the renewed FRE (*SSFRE*) on the contrary outperforms these state-of-the-art RF learning methods, which clearly demonstrates the efficiency of our semantic integration scheme.

5.2 Linguistic adjustment in KIBR relevance feedback

As shown in Sect. 4.2, for each region, each index keyword has its weight $P(K)$. In our Bayesian reasoning framework, we define $P(K)$ of keyword K as how well this keyword can reveal the user's search semantics in the current retrieval task. It is integrated into both QBE and QBK Bayesian reasoning for semantic-level RF.

Initially, all of the weights (probabilities) $P(K)$ are assigned a equal value. After the user provides RF images to our system, this $P(K)$ is adjusted to reflect the ability of each keyword to reveal semantics.

Suppose $P_1, P_2, \ldots, P_{n1}$ and $N_1, N_2, \ldots, N_{n2}$ denote the positive and negative examples respectively, and $P(K_1), P(K_2), \ldots, P(K_m)$ are the weights of m keywords. The RF learning procedure can be described as follows (Table 5):

Equations 21 and 22 can be viewed as positive/negative similarity accumulation based on user feedback and the *visual dictionary*, in which ground-truth labeling (visual dictionary) is combined with user supervision (positive/negative examples) in an exponential style for result re-ranking. With user engagement, keywords that are relevant to the user's retrieval concept would gain more attention, while irrelevant or contradicted keywords would lighten their impact in our Bayesian reasoning framework.

6 Experimental results and discussion

Our experiments were conducted on two evaluation databases. The first one is the COREL database (10,000 images), which shows the efficiency of our method by

Table 5 Keyword annotation adjustment in KIBR

```
For each keyword in the keyword set
  Set $PosSim = NegSim = 0$, which respectively mean the sum of the positive/negative similarity of this keyword;
      For each positive feedback image
          For each region in this image
              Select the most similar region in our pre-labeled set with this keyword, update the PosSim as:

$\mathrm{PosSim} = \mathrm{PosSim} + e^{-\|R_{\mathrm{Label}} - R_{\mathrm{Pos}}\|_2}$     (21)

    For each negative feedback image
      For each region in this image
          Select the most similar region in our pre-labeled set with this keyword, update the NegSim as:

$\mathrm{NegSim} = \mathrm{NegSim} + e^{-\|R_{\mathrm{Label}} - R_{\mathrm{Neg}}\|_2}$     (22)

  Set the weight $P(K)$ to PosSim/NegSim;
```

Table 6 Sample images from Flickr database and their keywords extracted from text descriptors

Image	Annotations	Image	Annotations
	Weather, storm, clouds, olivedowns, dam, groundtank, easter, girls, girlsdayout, desertlife, desert, outback		Bear,money, cash, fake
	Israel, eilat, lion, fish, red, sea		Flore, Northampton, canal, bridge, reflection

comparison with state-of-the-art algorithms; the second one is the Flickr database, which shows the efficiency in a real-world application.

The COREL 10,000: This database contains over 10,000 general purpose images based on 100 CD-ROMs published by COREL Corporation. Typically, each COREL CD-ROM of about 100 images represents one distinct category. The whole *COREL 10,000* was used for retrieval performance evaluation, while part of it (containing 2,000 images) was selected for annotation result evaluation, including autumn, bald eagles, bears, churches, coasts, elephants, exotic cars, fields, fighter jets, fish, fungi, helicopters, icebergs, pyramids, roses, sailboats, sunrises, tigers, trains, and waves. To construct the COREL query set (*COREL Query*), eight images of each above class were randomly selected from each category to form a query collection with 160 images.

The Flickr image database: This database contains 25,917 images downloaded from www.flickr.com together with their textual descriptors in an HTML page using the Flickr API. We adopted the class labels of the *COREL 10,000* (100 keywords) as initial keyword seeds that were sent to the Flickr API for downloading. We downloaded over 25,000 images from the Flickr online photo sharing database, averaging 250 images for each class. The annotation labels were extracted using the TF-IDF rule. Table 6 shows some examples of the Flickr image database and their keywords.

For the validation of the annotation performance, our annotation set was within the COREL 2,000, in which 92 images out of each category (autumn, bald eagles, bears, churches, coasts, elephants, exotic cars, fields, fighter jets, fish, fungi, helicopters, icebergs, pyramids, roses, sailboats, sunrises, tigers, trains, and waves) were selected to form the annotation set. We segmented photos within this annotation set and manually annotated each region of this set, which was used for training PWC SVM classifiers. The rest of this COREL 2,000 dataset (8 × 10 images) was left for validation of annotation performance. For retrieval performance, the *COREL 10,000* and *Flickr 25,000* data sets were used.

All our experiments were performed on a 2.5 GHz Intel® Celeron® machine with 512 MB RAM. The algorithms were implemented using C++ in a Microsoft Visual Studio environment. We used *Precision* and *Recall* to evaluate retrieval performance:

$$\mathrm{Precision} = \frac{n_{\mathrm{relevant}}}{n_{\mathrm{return}}} \tag{23}$$

$$\mathrm{Recall} = \frac{n_{\mathrm{relevant}}}{n_0} \tag{24}$$

Table 7 Example results of annotation propagation in database regions

Keyword	Total appearance					Correct appearance				
	1	2	3	4	5	1	2	3	4	5
Sun	27	35	9	4	2	20	22	4	0	0
Sky	32	28	19	11	54	16	13	14	44	27
Water	26	19	8	8	1	22	13	4	3	0
Bear	32	21	16	12	5	20	10	6	4	1
Tree	16	14	15	20	20	11	10	11	7	10
Grass	18	18	12	20	17	15	13	8	10	7
Building	27	16	23	26	5	21	12	10	0	2
Mountain	13	28	66	29	38	7	10	25	18	13
Elephant	22	25	12	7	3	16	11	5	3	1
Tiger	4	10	7	5	4	4	4	3	3	0
Wolf	19	21	36	18	41	9	9	9	13	19
Snow	12	10	16	9	50	7	3	5	6	18

n_{relevant} represents the number of related images in the returned images collection; n_{return} represents the number of images returned from each retrieval process; n_0 is the number of images in each image class (In our experiments 100). We used the *standard deviation* of either Precision or Recall to measure the robustness of the corresponding retrieval algorithm.

We briefly present the visual features we used as follows, in which (1) was used in clustering-based image segmentation and (2)–(4) were used in region representation as well as RF:

(1) Region-level features: as presented in Sect. 2, we extracted a six-dimensional color and textual mixture feature for each region

(2) Color features: in our system, the accumulated color histogram of the H component (HSV color space) was adopted as the color features. It was quantized into 64 bins to enhance computation efficiency, and a smoothing operation was applied to this quantized histogram to filter image noise.

(3) Texture features: the texture co-occurrence matrix was calculated as a texture descriptor; it is popular for texture representation [9].

(4) Auto-Correlogram [10]: the auto-Correlogram, which combines spatial information with the color histogram, has been proven to be very effective in CBIR [13]. Similar to [13], we quantized RGB color space into 64 bins and the pixel correlations within distances 1, 3, 5, and 7 were extracted and calculated to form a 256-dimensional feature.

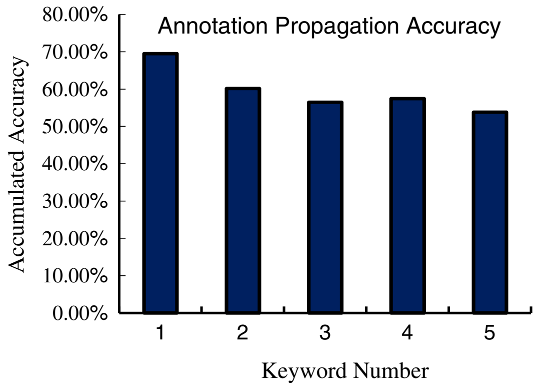


Fig. 4 Accumulated annotation accuracy

6.1 Annotation propagation performance

First, we tested our annotation propagation algorithm over our 2,000 COREL subset, and compared our annotation propagation results with the manual labeling results. To construct the training set, in each region a single keyword was manually labeled to capture its linguistic semantic meaning. To address the sample asymmetry and sample insufficiency issues in keyword classifier learning, inter-keyword (inter-class) SVM classifiers were trained by bagging training samples. Table 7 presents the annotation accuracy in annotation positions 1–5 of 12 randomly selected keywords in our visual dictionary. Table 10 shows the accumulated annotation accuracy of the top 1, 2, 3, 4, and 5 keywords of the first 12 keywords in our visual dictionary. The overall accumulated annotation accuracy (Eq. 25) of our visual dictionary is demonstrated in Fig. 4.

Table 8 Example results of accumulated annotation accuracy in all regions of the database

Keyword	Annotation accuracy (first–fifth keywords) (%)				
Sun	74.07	67.74	64.79	61.33	59.74
Sky	50.00	48.33	54.43	96.67	79.17
Water	84.62	77.78	73.58	68.85	67.74
Bear	62.50	56.60	52.17	49.38	47.67
Tree	68.75	70.00	71.11	60.00	57.65
Grass	83.33	77.78	75.00	67.65	62.35
Building	77.78	76.74	65.15	46.74	46.39
Mountain	53.85	41.46	39.25	44.12	41.95
Elephant	72.73	57.45	54.24	53.03	52.17
Tiger	100.0	57.14	52.38	53.85	46.67
Wolf	47.37	45.00	35.53	42.55	43.70
Snow	58.33	45.45	39.47	44.68	40.21
Total	69.44	60.12	56.43	57.40	53.78

Table 9 Annotation precision and recall of partial visual dictionary in all database regions

Keyword	Total app	Correct app	Suppose app	Precision	Recall
Sun	73	46	51	0.630	0.902
Sky	217	114	123	0.525	0.927
Water	62	46	47	0.742	0.979
Bear	86	41	48	0.477	0.854
Tree	76	49	68	0.645	0.721
Grass	86	53	65	0.616	0.815
Building	93	45	66	0.484	0.682
Mountain	174	73	81	0.420	0.901
Elephant	69	36	39	0.522	0.923
Tiger	30	14	18	0.4667	0.778
Wolf	135	56	65	0.415	0.862
Snow	97	39	45	0.402	0.867

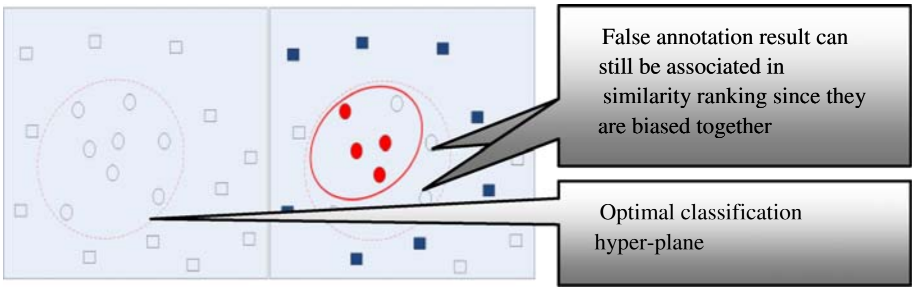
Table 10 Probability of keyword similarities and its modification using RF learning

Keyword similarity probability and its evolution					
Original			Feedback Result		
Top5	Keyword	Prob.	Top5	Keyword	Prob.
1	Flower	0.27	1	Flower	0.69
2	Leaf	0.26	2	Grass	0.53
3	Land	0.19	3	Leaf	0.51
4	Train	0.13	4	Fungi	0.33
5	Building	0.09	5	Forest	0.21

$$\text{Accumulated Accuracy}_{\text{Topi}} = \frac{\text{Correct Annotation of Current Keyword}}{\text{First i Annotated Keywords in an image}} \tag{25}$$

As seen from Tables 7–9, some keywords were well annotated (tiger, water, grass) while some others did not do so well (sky, wolf, mountain). This was because the inter-keyword SVM trained for annotation propagation would shrink toward the more compact class when the compactness of two classes' visual features was unequal. For example, the PWC SVM between keyword "mountain" and keyword "water" definitely shrank toward "water" since the visual features of water are more uniform than those of "mountain" (The regions of "water" are more visually uniform while the

Fig. 5 False annotations are biased together, and hence are still associated in similarity ranking



False annotation result can still be associated in similarity ranking since they are biased together

Optimal classification hyper-plane

regions of "mountain" may be at visually different, with different covers, at different seasons, etc). Hence, the resulting SVM would be biased away from the true "virtual" separation hyper-plane between "mountain" and "water" toward the vectors of "water". However, this bias is made synchronously for all regions, hence the calculation results of their corresponding semantic factors would remain high since their keyword annotations are biased together; this was demonstrated in retrieval experiments (Sect. 6.3) (Fig. 5).

As presented in Table 8, the accumulated annotation accuracy decreased as the number of annotated keywords in each region increased from 1 to 5. This is inevitable since the keywords that actually need to be annotated to each region are limited. Hence our soft annotation strategy would unavoidably bring in some "reluctant" keywords for some regions. However, the accumulated accuracy of our soft annotation results remained over 50% in the top 5 and 60% in the top 2. In spite of the unavoidable reluctant keywords, our annotation propagation algorithm still demonstrated a good semantic description capability. Furthermore, this "soft" semantic description is extremely suitable in the case that the segmentation result is unsatisfactory and groups multiple objects in a single region. In such cases, these reluctant segmentation regions should be of course be labeled with multiple keywords. Moreover, in some mis-annotated instances, the confidence factor of each keyword that indicates its corresponding confidence level would be adjusted in RF learning. Hence such mis-annotated information is diluted.

6.2 Semantically supervised retrieval by visual and textual fusion

We define *region retrieval* as retrieving image similarities based on region similarity. First, a user-provided query image is annotated using the trained PWC SVM bagging network, and the users interact with our system to give evaluations about the quality of the query example annotations (well annotated or poorly annotated). Using the user's context (annotation judgment) and content (query image), retrieval is achieved by fusing the visual and textual cues in similarity ranking, with different choices (SSIRM or KIBR) deployed for different annotation qualities.

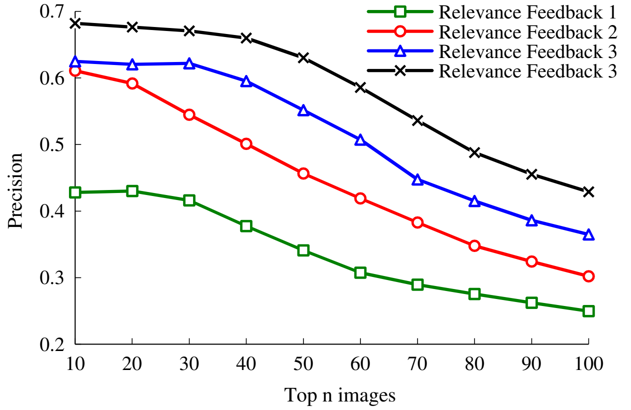


Fig. 6 Precision of semantic-supervised integrated region matching (SSIRM)

In SSIRM, region retrieval concentrates on the entire image level, while in KIBR, region retrieval refers to not only searching for similar images but also measuring the image similarity based on the target region. Also, the matching degree of the query image and database image are leveraged into our similarity ranking. In SSIRM, the semantically supervised IRM distance is calculated to measure the similarity between each image in the database and the query image; in KIBR, the image similarity is measured using probabilistic Bayesian reasoning, in which *query-by-example* in CBIR can be regarded as a process to deduce the conditional probability $P(I|I_Q)$ of each image I in the database given query example I_Q.

To test system performance with the state-of-the-arts in a large-scale data environment, we tested our 160-image query set on the *COREL 10,000*, in which our annotation algorithm propagated multi-labels onto the 2,000 image subset as mentioned before (this subset contains all topics of our query set). We evaluated the performance improvement of semantic-supervised region retrieval compared with retrieval algorithms using visual features at both the image global-level and region-level. The retrieval results (precision and recall) of our SSIRM are presented in Figs. 6 and 7. Figure 8 compares the region retrieval results of our method (SSIRM) and IRM. As shown in Fig. 8 the precision of our SSIRM obviously outperformed IRM, which demonstrates the efficiency of semantic supervision.

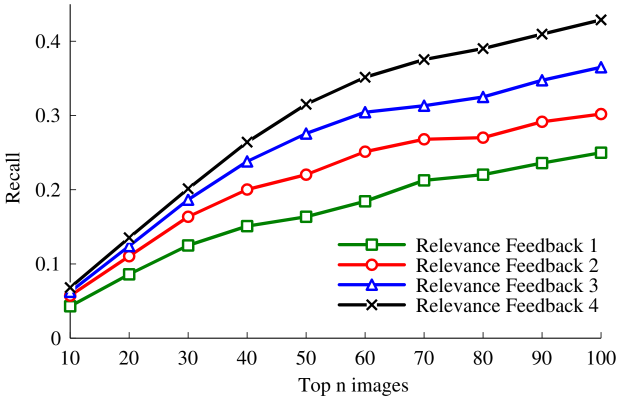


Fig. 7 Recall of semantic-supervised integrated region matching (SSIRM)

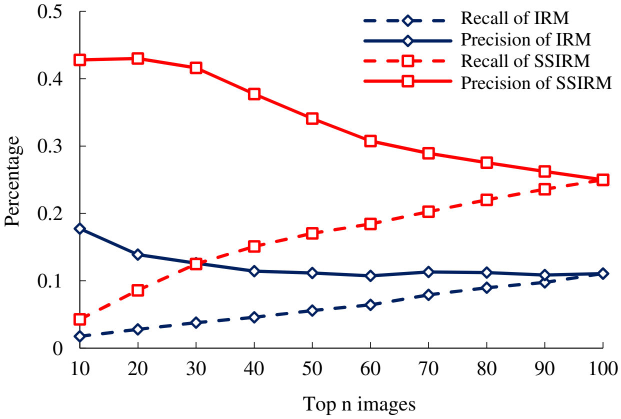


Fig. 8 Performance comparison of IRM and SSIRM in initial query

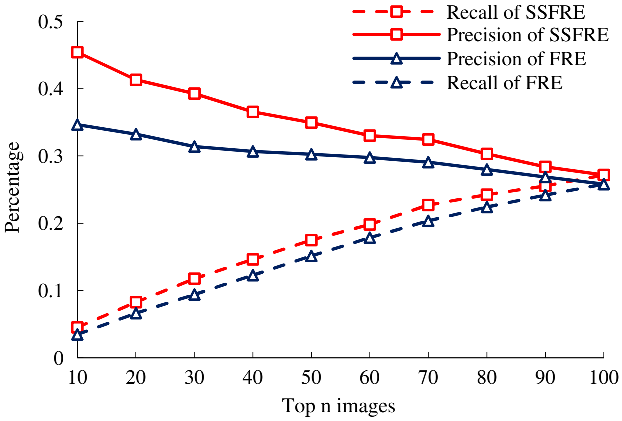


Fig. 9 Initial query: visual feature versus visual and textual feature

Figure 9 shows the performance enhancement based on our semantic supervision method using global features (Color Histogram + Text Co-Occurrence + Auto-Correlogram), which produced better results than FRE using solely visual features.

Although region-based retrieval methods are more semantically sensitive to human image perception, the patch test results of region retrieval are lower than those of traditional global feature matching methods in the COREL image database. This is because intra-class similarities are higher than inter-class similarities in the COREL image database, and the objects contained in the images have diverse visual features. In fact, the same semantic objects may reveal diverse visual features. However, it should be emphasized that region-retrieval is definitely perceptually closer to human image perception.

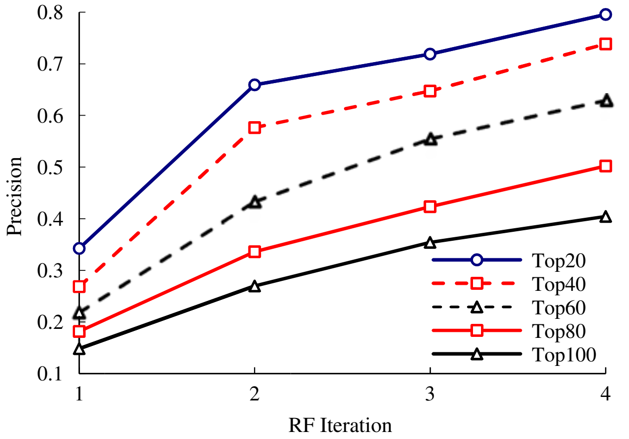


Fig. 10 KIBR precision enhancement for the top 20–100 returned images

6.3 Semantically supervised relevance feedback learning

Finally, the fusion efficiency of our strategy was demonstrated by comparison with three RF learning strategies in the *COREL 10,000*, including: (1) a classical FRE RF learning strategy [26], SVM RF learning [20]; (2–3) two state-of-the-art methods: active Learning SVM [21], and asymmetry bagging SVM [19]. It has been demonstrated that these state-of -the-art RF methods definitely outperform FRE by a large margin in visual-only scenarios. However, by integrating our semantic RF learning, the renewed FRE (SSFRE) on the contrary outperformed these state-of-art RF learning methods, demonstrating the efficiency of our proposed method.

Figures 10 and 11 show the enhancement of KIBR performance using our RF learning method (Sect. 5.2). It is obvious that after three RF operations, the average retrieval precision enhancement was over 35% in the case that $n < 60$; the average retrieval recall enhancement was near 25% when $n > 80$.

The efficiency of the proposed SSIRM+SVM method was compared with IRM [23] + SVM to demonstrate the efficiency of semantic supervision. Figures 12, 13, 14 and 15 show the RF performance enhancement of SSIRM over SRM+SVM, which was implemented as the baseline algorithm for region-based retrieval and RF learning. Superior performance improvement was demonstrated from the

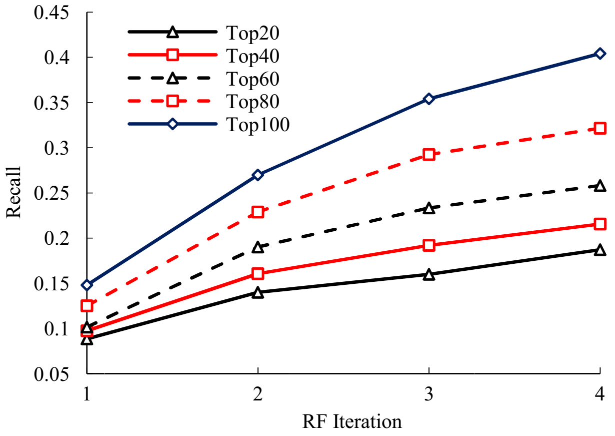


Fig. 11 KIBR recall enhancement for the top 20–100 returned image

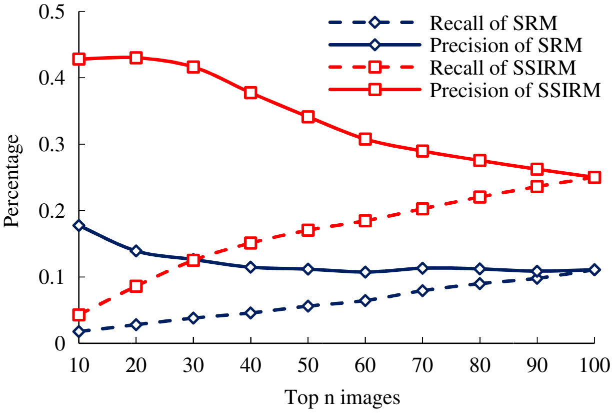


Fig. 12 SRM versus SSIRM at initial retrieval

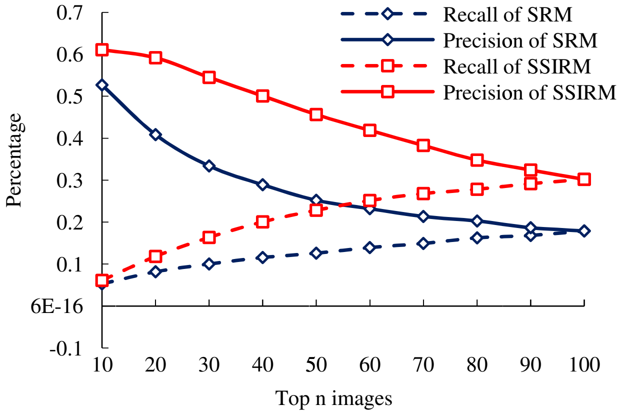


Fig. 13 SRM versus SSIRM at first RF

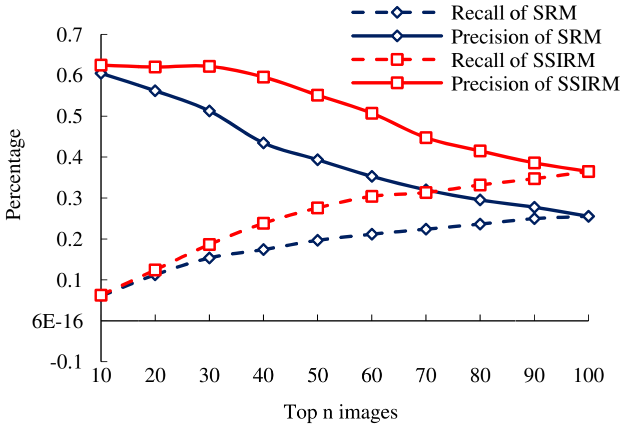


Fig. 14 SRM versus SSIRM at second RF

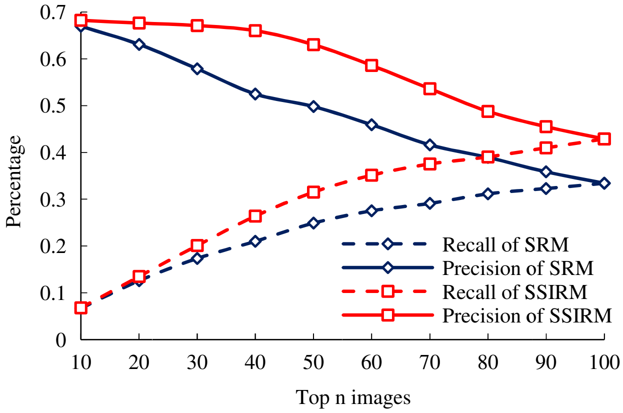


Fig. 15 SRM versus SSIRM at third RF

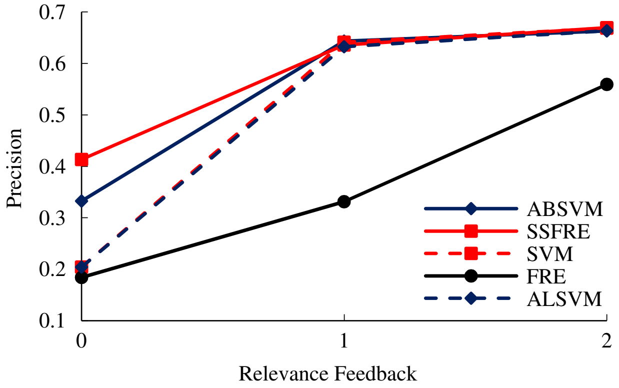


Fig. 16 RF learning precision in top 20

initial retrieval to the third RF learning. The average precision and recall enhancements in the top 100 images were about 10% during RF learning in each round.

Figures 16, 17, 18, 19, 20, 21, 22, 23, 24 and 25 present the RF learning comparison of our method with classical and state-of-art methods in the global feature space. These methods included: FRE [26], SVM [20], active learning SVM (ALSVM) [21] and asymmetry bagging SVM (ABSVM) [19] RF learning methods. Although only a simple FRE method was integrated with our semantic RF learning, it was obvious that, during three RF iterations, our semantic integration strategy gained significant performance enhancements in both precision rates and recall rates among the top 20–100 images. Moreover, since we conducted only FRE learning at the visual level, the efficiency of the semantic RF learning

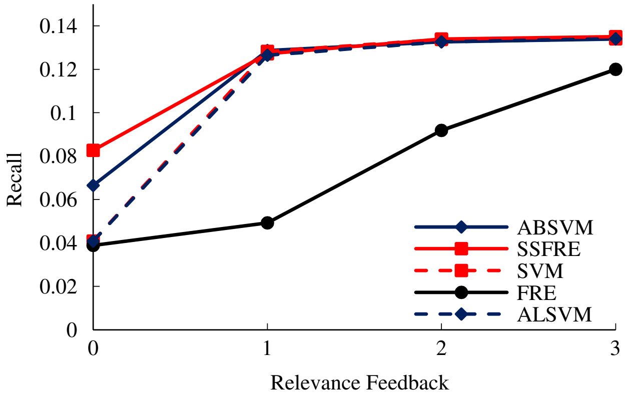


Fig. 17 RF learning recall in top 20

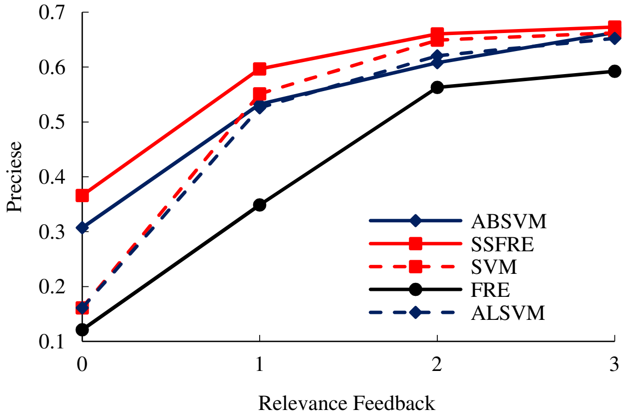


Fig. 18 RF learning precision in top 40

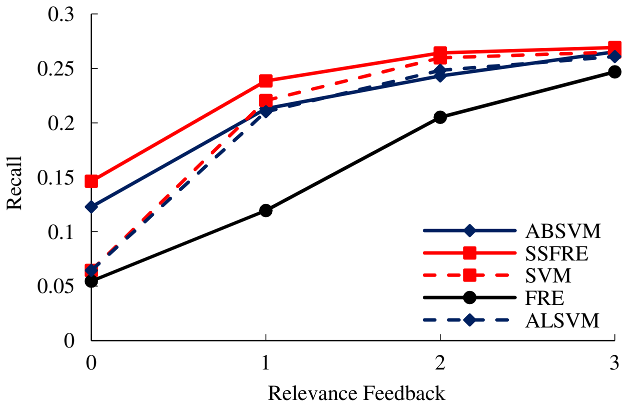


Fig. 19 RF learning recall in top 40

is evidently demonstrated. To ensure valid results, we first re-implemented ABSVM, SVM, and ALSVM as in the original papers to achieve comparable results in their datasets. Then we leveraged the tuned parameters of these classifiers in RF learning on our validation dataset. The reason behind such improvement is the integration of semantic supervision, since the FRE itself indeed could not produce results comparable to SVM and its improvements.

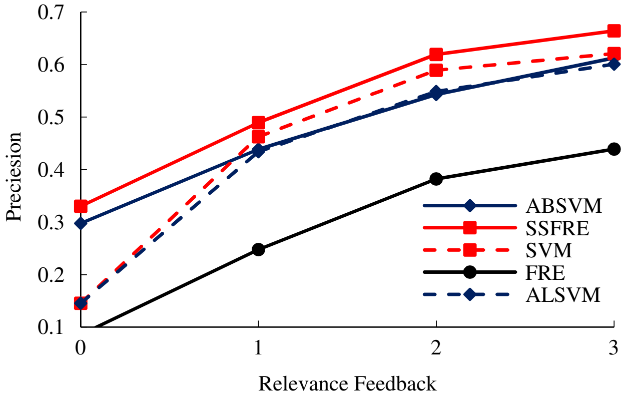


Fig. 20 RF learning precision in top 60

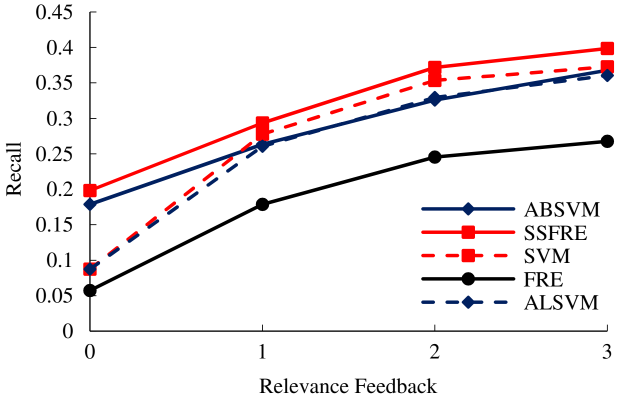


Fig. 21 RF learning recall in top 60

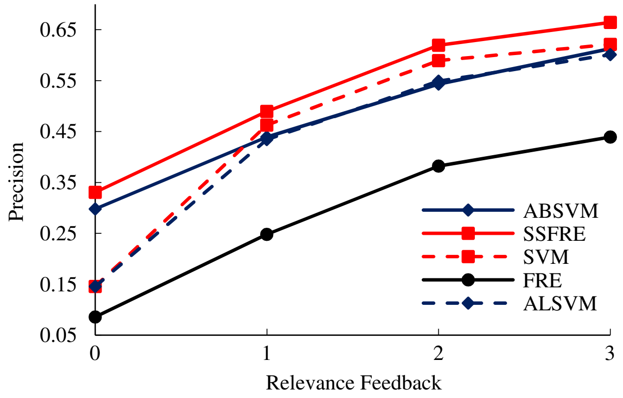


Fig. 22 RF learning precision in top 80

We present a case study for the KIBR: Fig. 26 shows the query image and its segmentation results. Table 10 shows in detail the keyword similarity and its adjustment after five RF operations. As demonstrated in Table 10, the keywords “flower” and “leaf” gain higher similarity probability after user feedback, which is exactly consistent with human image perception.

Finally, the robustness of our proposed method was demonstrated by the retrieval *standard deviation*curves on our

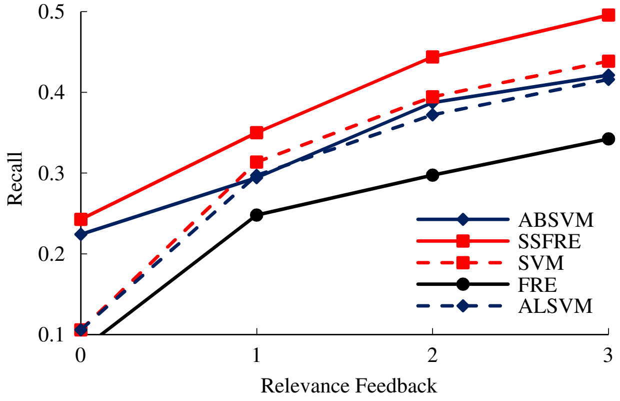


Fig. 23 RF learning recall in top 80

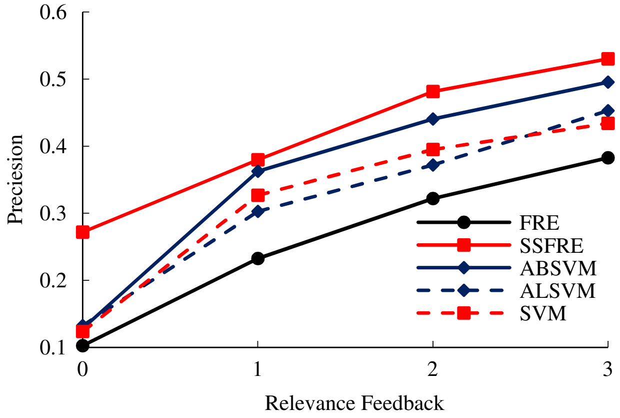


Fig. 24 RF learning precision in top 100

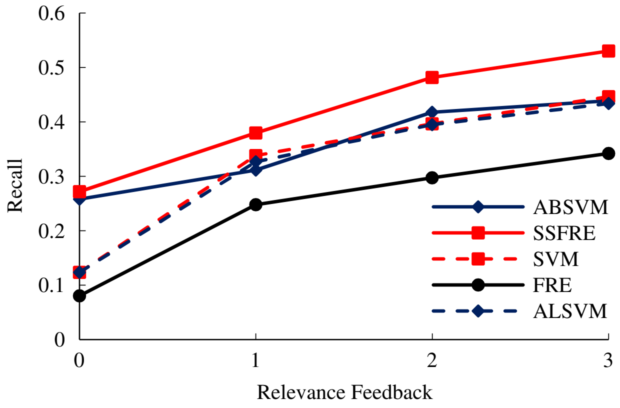


Fig. 25 RF learning recall in top 100

COREL 2,000 database (not in COREL 10,000 this time). Figures 27, 28 and 29 present the *standard deviation* comparison of both *SSIRM* and *KIBR* with SVM in the first, second and third RF learning processes. When the number of returned images was more than 40, our SSIRM and KIBR gave smaller values than SVM. In other words, both *KIBR* and *SSIRM* are more stable than an SVM-based RF learning scheme.

It is worth mentioning that, along with the algorithms we present, the most important contribution of this paper is the effective annotation and retrieval reinforcement framework, which makes image retrieval closer to human perception. By utilizing other new efficient or sophisticated strategies to replace the visual-level ranking and learning of this framework, our system performance could be further enhanced.

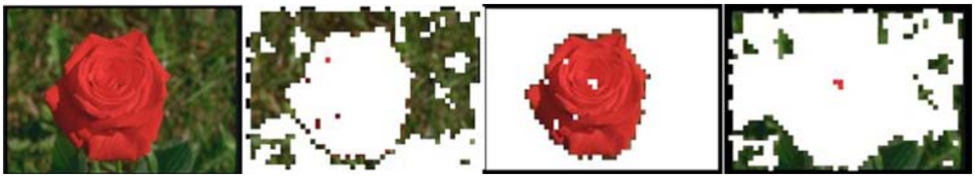
6.4 Performance evaluation on Flickr database

To testify the effectiveness of our proposed visual and textual fusion framework in a real-world scenario, the Flickr web database was also used for performance demonstration. A randomly selected query set (*Flickr query*) was constructed, containing 160 photos from 20 classes, which was similar to our query set construction in the COREL database. The precision evaluation was defined as the ratio between the number of returned images belonging to the class of the query image and the total number of returned images. We treated user labeling as the keyword annotation for each image region (in such a phase, the labels were abundant and over-fitting for each region, but this abundance in "semantic factor" could be easily diluted in the subsequent SSIRM process); on this basis the multi-label assumption was naturally achieved over the whole database.

We present two groups of experiments on this database. First, we evaluated the system performance using the *Flickr query* set, based on which six rounds of RF were conducted. This was leveraged to evaluate the real-world query in a real-world database. Second, we evaluated the system performance using the *COREL query* set, based on which six rounds of RF were also conducted. This is because we are interested in viewing the differences and variations between the COREL and Flickr databases.

As presented in Fig. 30, with the *Flickr Query* set the initial precision of our visual and textual fusion system was

Fig. 26 A query image and its segmentation results



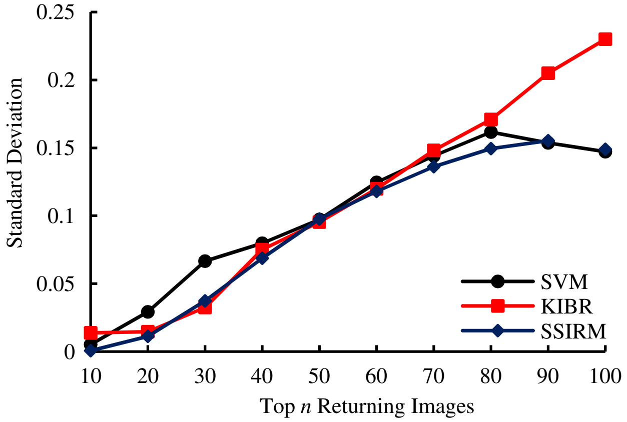


Fig. 27 Retrieval standard deviation in first RF

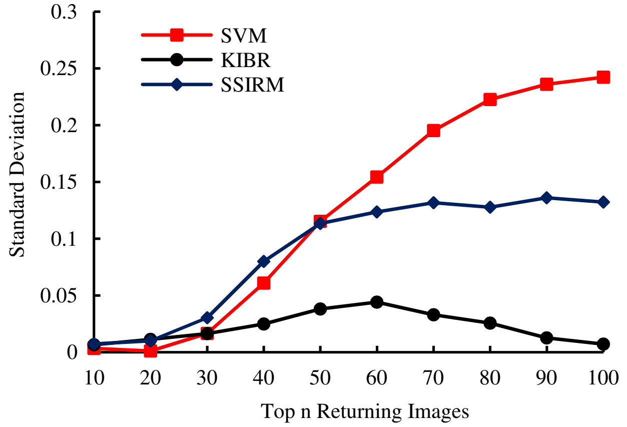


Fig. 28 Retrieval standard deviation in second RF

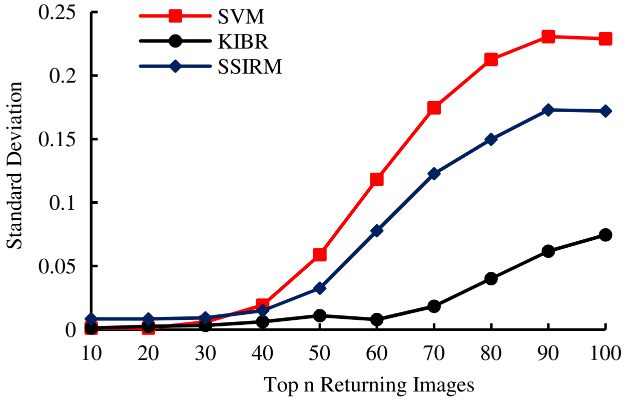


Fig. 29 Standard deviation in third RF

over 25% in the top 10-top 50, and over 20% in the top 60-top 90. For the top 10 returned results: after the 1st RF, the precision enhancement was over 10%; after the 2nd RF, it was over 6%, achieving 45% in the top 10. Also, after the 5th RF, the precision was over 50% in the top 10-top 30, and over 35% in the top 80. Figure 31 shows a somewhat astonishing phenomenon: Using the *COREL Query* set, the performance was slightly worse than the results in Fig. 30. In our opinion,

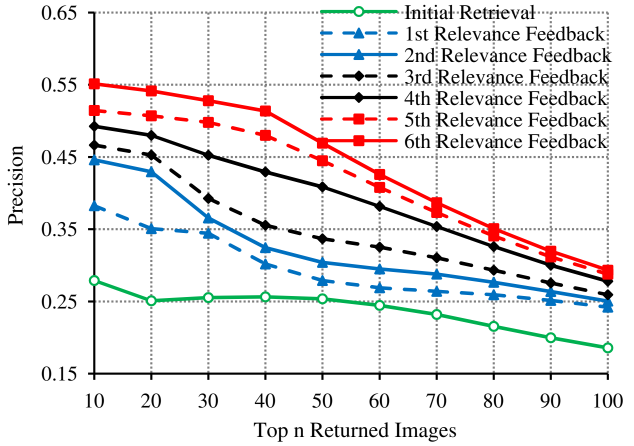


Fig. 30 Retrieval precision on the Flickr database using the Flickr query set

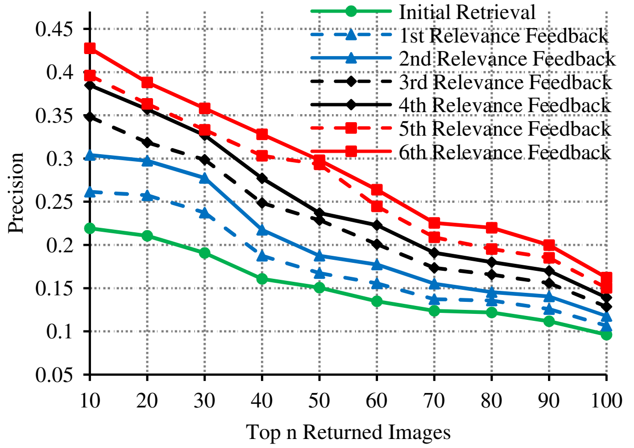


Fig. 31 Retrieval precision on the Flickr database using the COREL query set

this indicates that the constitution of the COREL database was indeed different from the web user's perspective given identical keywords. Meanwhile, the initial ranking results show that although the top 10-top 30 precision were almost similar between Figs. 30 and 31, the subsequent precision (top 40-top 100) of the *Flickr Query* was better than that of the *COREL Query*, which further indicates the dissimilarity between the COREL database and the Flickr database. Also, comparing Fig. 30 with earlier results in Figs. 7, 9, and 12, the web-based dataset was far more complicated than the COREL database.

7 Conclusion

In this paper, we focus on fusion of region annotation with region-based image search, which is a promising solution to address the semantic gap in traditional CBIR. We present a unified framework for semantically supervised region retrieval, which also offers a novel annotation- interaction-retrieval interaction pattern for users to precisely express their retrieval target.

In offline region annotation learning, we present a PWC-SVM bagging network to cope with the problems of sample insufficiency and sample asymmetry. In semantically supervised image retrieval, we explore the fusion of visual and textual information from both soft reasoning (if the user judges the query image to be well annotated) and Bayesian reasoning (if the user judges the query image to be poorly annotated) aspects. In particular, SSIRM can largely reduce the negative effect of imprecise segmentation by soft matching, and KIBR supports both QBK and QBE interfaces, both with solid theoretical foundations and a natural integration of the *visual dictionary* from annotation.

In experimental validation, first we compared our framework with state-of-the-art methods on the *COREL 10,000*. Then we evaluated our system performance over a Flickr 25,000 web image database. As demonstrated in our experiments, by annotation and retrieval reinforcement our fusion framework and annotation-interaction-retrieval strategy can considerably shorten the semantic gap in image retrieval. However, there are still open problems to be addressed in our future work. The propagation of our keyword annotation only needs the training of pair-wise coupling SVMs between two keywords. A simplifying assumption is made here: keywords are independent in semantic meaning. However, the linguistic meanings of words are often related (For example, “elephant” belongs to “mammal”). Our future research will concern this problem.

Acknowledgements This work was supported by State 863 High Technology R&D Project of China (No. 2006AA01Z197); National Key Basic Research Program of China (No.2009CB320905), Natural Science Foundation of China (No. 60472043).

References

1. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Region-based image querying. Proceeding of IEEE Workshop on Content-Based Access of Image and Video Libraries, pp. 42–49 (1997)
2. Chang, E., Goh, K., Sychay, G., Wu, G.: CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machines. IEEE Trans. Circuits Syst. Video Technol. **13**(1), 26–38 (2003)
3. Chen, Y., Wang, J.Z.: A region-based soft feature matching approach to content-based image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **24**(9), 1252–1267 (2002)
4. Cox, I.J., Miller, M.L., Minka, T.P., Papathomas, T.V., Yianilos, P.N.: The Bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments. IEEE Trans. Image Process. **9**(1), 20–37 (2000)
5. Datta, R., Ge, W., Li, J., Wang, J.Z.: Toward bridging the annotation-retrieval gap in image search. IEEE Multimed. **14**(3), 24–35 (2007)
6. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. **40**(2), 1–60 (2008)
7. Duygulu, P., Barnard, K., de Freitas, N., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. Eur. Conf. Comput. Vis. **4**, 97–112 (2002)
8. Goh, K.-S., Chang, E.Y., Li, B.: Using one-class and two-class SVMs for multiclass image annotation. IEEE Trans. Knowl. Data Eng. **17**(10), 1333–1346 (2005)
9. Haralick, R.M., Shanmugam, K., Dinstein, I.: Texture features for image classification. IEEE Trans. Syst. Man Cybern. **3**, 610–621 (1973)
10. Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J., Zabin, R.: Image indexing using color correlogram. IEEE Int. Conf. Comput. Vis. Pattern Recognit. pp. 762–768 (1997)
11. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 119–126 (2003)
12. Ji, R., Lang, X., Yao, H., Zhang, Z.: Semantic sensitive region retrieval using keyword integrated Bayesian reasoning. Int. J. Innov. Comput. Inf. Control **3**(6), 1645–1656 (2007)
13. Jing, F., Li, M., Zhang, H.-J., Zhang, B.: A unified framework for image retrieval using keyword and visual features. IEEE Trans. Image Process. **14**(7), 979–989 (2000)
14. Lavrenko, R.M.V., Jeon, J.: A model for learning the semantics of pictures. Annual Conference on Neural Information Processing Systems (2003)
15. Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1075–1088 (2003)
16. Lu, Y., Zhang, H.-J., Liu, w., Hu, C.: Joint semantic and feature based image retrieval using relevance feedback. IEEE Trans. Multimed. **5**(3), 339–347 (2003)
17. Rahman, M.M., Bhattacharya, P., Desai, B.C.: A framework for medical image retrieval using machine learning & statistical similarity matching techniques with relevance feedback. IEEE Trans. Inf. Technol. Biomed. **11**(1), 58–69 (2007)
18. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1349–1380 (2000)
19. Tao, D., Tang, X., Li, X., Wu, X.: Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **28**(7), 1088–1099 (2006)
20. Tong, S., Chang, E.: Support vector machine active learning for image retrieval. Proceeding of ACM International Conference on Multimedia, pp. 107–118 (2001)
21. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
22. Veltkamp, R.C., Tanase, M.: Content-based image retrieval systems: A survey, technical report UU-CS-2000–34, Department of Computing Science, Utrecht University (2000)
23. Wang, J.Z., Li, J., Wiederhold, G.: SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. IEEE Trans. Pattern Anal. Mach. Intell. **23**(9), 947–963 (2001)
24. Wang, T., Yong, R., Sun, J.-G.: Constraint based region matching for image retrieval. Int. J. Comput. Vis. **56**(1/2/3), 37–45 (2004)
25. Wu, T., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Int. J. Mach. Learn. Res. **10**(5), 975–1005 (2004)
26. Yong, R., Huang, T.S., Mehrotra, S., Ortega, M.: Relevance feedback: A power tool for interactive content-based image retrieval. IEEE Trans. Circuits Syst. Video Technol. **8**(5), 644–655 (1998)
27. Zhang, R., Zhang, Z.: Hidden semantic concept discovery in region based image retrieval. Comput. Vis. Pattern Recognit. **2**, 996–1001 (2004)